The Check Float Puzzle

Jeffrey M. Lacker

Although the last few years have seen a dramatic surge in interest in new electronic payment instruments, consumers and businesses in the United States still write checks in vast numbers. Nearly 63 billion checks were written in 1995 according to one estimate, representing 78.6 percent of all noncash payments (Committee on Payment and Settlement Systems of the central banks of the Group of Ten countries 1995). Check use has continued to expand in recent years, despite the increased use of debit cards and the automated clearinghouse; the per capita number of checks written grew at an average annual rate of 1.3 percent from 1991 to 1995. Moreover, forecasts call for check use to remain around current levels for the foreseeable future (Humphrey 1996). Because the social costs associated with the use of paper checks constitutes the majority of the real resource costs of the payment system—65.4 percent according to David Humphrey and Allen Berger (1990)—it will be important to continue to seek improvements in the efficiency of the check system in the years ahead.

The efficiency of check clearing is affected by the arrangements governing presentment and payment. These arrangements have a feature that is, for economists, puzzling. Helen writes a check to John for, say, $100. When the check is ultimately presented to Helen's bank for payment, the bank pays $100, and deducts $100 from Helen's account. What is surprising, from an economist's point of view, is that the bank pays the same amount, $100, *no matter how long it took for the check to be presented*. This implies that John's bank earns an additional day's interest by getting the check to Helen's bank one day sooner. This feature is puzzling because it is difficult to identify any significant social benefits to Helen or Helen's bank from getting a check from John's bank one day sooner; certainly nothing approaching the magnitude of one day's interest.

Helen Upton deserves grateful thanks for research assistance. Gayle Brett, Andreas Hornstein, Tom Humphrey, Ned Prescott, Marsha Shuler, and John Weinberg provided helpful comments on an earlier draft, but the author remains solely responsible for the contents of this article. The views expressed do not necessarily reflect those of the Federal Reserve Bank of Richmond or the Federal Reserve System.

Check float is the time between when a check is tendered in payment and when usable funds are made available to the payee (John in our example).[1] Because John and his bank bear the opportunity cost of foregone interest until the check is presented, they have an incentive to minimize the float. But check float provides interest income for Helen and her bank. Under current arrangements Helen and her bank implicitly reward John and his bank for reducing check float. Helen's bank stands ready to turn over their float earnings. John's bank thus has an incentive to capture those float earnings by accelerating presentment. Another way to state the puzzle is that the benefits to Helen and her bank do not seem to justify the incentive provided to John and his bank to minimize check float. For this reason I call it the "check float puzzle."

The resolution of this puzzle is of more than intellectual interest. Because collecting banks forgo interest earnings on the checks in their possession, they have a strong incentive to present them as quickly as possible in order to minimize the interest foregone. Collecting banks are motivated to incur significant real resource costs to accelerate the presentment of checks. Check processors, including the Federal Reserve Banks, routinely compare the cost of accelerating presentment to the value of the float. Checks are sorted at night and rapidly shipped across the country. But if there is little or no social benefit of accelerating the presentment of checks, then much of the real resource costs associated with check processing and transportation would represent waste from the point of view of the economy as a whole. It may be possible to alter this puzzling arrangement and improve the efficiency of the payment system.

The check float puzzle can be directly attributed to the fact that the laws and regulations governing check clearing mandate *par presentment*; the payor owes the face value of the check, no matter when the check arrives. Par presentment implies that the real present discounted value of the proceeds of clearing the check are larger the faster the check is presented. Par presentment essentially fixes the relative monetary rewards to alternative methods of clearing, taxing slower methods of clearing relative to faster methods. As with any regulation that fixes relative prices, there is the potential to distort resource allocations. In this article I argue that the distortion appears to be significant. This is only part of the story, however. There could be offsetting benefits that make par presentment a good thing. To justify current arrangements there would have to be social benefits of clearing checks quickly that payees and their banks—the ones deciding how fast to clear the check—do not take into account.

The check float puzzle is of interest to the Federal Reserve System (the Fed), both as payment system regulator and as the largest processor of checks. In the 1970s the Federal Reserve Banks established a number of Remote Check

[1] This use of the word float follows Humphrey and Berger (1990, p. 51). The reader should be aware that some writers use the term float in a narrow sense to refer to the time between when the payee is credited and the payor is debited: see, for example, Veale and Price (1994).

Processing Centers (RCPCs) around the country with the avowed goal of accelerating the presentment of checks (Board of Governors of the Federal Reserve System 1971; Board of Governors of the Federal Reserve System 1972). Critics have argued recently that Federal Reserve operations should be consolidated to take advantage of economies of scale in check sorting (Benston and Humphrey 1997). But closing down Fed offices could increase the amount of time it takes to collect some checks. Should this result be counted against the decision to close an office? More generally, when performing a cost-benefit analysis of alternative payment system arrangements, what value should be placed on changes in the speed of check collection?

Check Float

A few words about how check clearing works will be useful as background. Checks provide a simple arrangement for making payments by transferring ownership of book-entry deposits. Helen (the "payor") writes a check and gives it to John (the "payee"). John deposits the check in his bank, which then initiates clearing and settlement of the obligation. A check is a type of financial instrument or contingent claim. It entitles the person or entity named on the check, the payee, to obtain monetary assets if the check is exchanged in accordance with the governing laws and regulations. One noteworthy feature of the check is that the holder of the check is entitled to choose when the check is exchanged for monetary assets. In other words, the check represents a demandable debt.

John's bank has a number of options available for getting the check to Helen's bank for *presentment*. John's bank could present directly, transporting the check itself or by courier to Helen's bank. Alternatively, the check could be presented through a *clearinghouse* arrangement in which a group of banks exchange checks at a central location. Another option is to send the check through a *correspondent bank* that presents the check in turn to Helen's bank. Or the check could be deposited with a Federal Reserve Bank, which then presents the check to Helen's bank. These intermediary institutions could themselves send the check through further intermediaries, such as clearinghouses, other correspondent banks, or other Reserve Banks.

The length of time it takes to present a check depends on where the check is going and on how John's bank decides to get it there. First, the checks received by John's bank during the business day are sorted based on their destination. Sorting generally occurs during the early evening hours. Afterward, many checks can be presented to the paying bank overnight. A check drawn on a nearby bank might be presented directly early the next morning. A group of neighboring banks that consistently present many checks to each other might find it convenient to organize a regular check exchange or clearinghouse in which all agree to accept presentment at a central location. Checks drawn on

local clearinghouse banks can generally be presented before the next business day.

For checks drawn on other nearby banks it might be advantageous to clear via a third party, such as a check courier, a correspondent bank, or the Federal Reserve. A third-party check processor posts a deadline, usually late in the evening, by which local checks must be deposited in order to be presented the next day. Third parties also clear checks drawn on distant banks. Often such checks can be presented by the next day as well, especially checks drawn on banks located in cities with convenient transportation links. For checks drawn on remote and distant locations, however, an additional day or two may be needed to get the check where it is going. For example, a check drawn on a bank in Birmingham, Alabama, and deposited at the Federal Reserve Bank of Richmond is usually presented to the Birmingham bank in one day, while a check drawn on a bank in Selma, Alabama, is usually presented in two days.

When does John's bank collect funds from Helen's bank? If the two banks do not have an explicit agreement providing otherwise, Helen's bank is obligated to pay John's bank on the day her bank receives the check, provided it is received before the appropriate cutoff time. If the check is presented by a Federal Reserve Bank, the cutoff time is 2:00 p.m.; if anyone else presents the check, the cutoff time is 8:00 a.m. Helen's bank is obligated to pay by transfer of account balances at a Reserve Bank or in currency; in practice Reserve Bank account balances are the rule. Checks presented after the cutoff are considered presented on the following business day.

A majority of the checks in the United States are presented in time for payment the next business day. According to a recent survey by the American Bankers Association (1994), over 80 percent of local checks are presented within one business day, while only about half of nonlocal checks are presented within one business day (Table 1). Over 90 percent of the dollar volume of checks cleared through the Federal Reserve are presented within one business day.

What's the Puzzle?

The puzzle is that *the paying bank pays the same nominal amount no matter how many days it takes to clear the check.* Helen's bank pays John's bank the face value of the check whether it takes one day, two days, or two weeks to clear. To put it another way, an outstanding check does not earn interest while the check is being cleared. The implication is that clearing a check one day faster allows the presenting bank to earn an extra day's interest. The presenting bank's gain is the paying bank's loss, however; Helen's bank gives up one day's interest. Why are arrangements structured this way?

At a superficial level the answer is transparent. The presentment of checks is governed by the Uniform Commercial Code, the Federal Reserve Act, and

Table 1 Number of Days It Takes to Receive Available Funds on Checks Deposited through Banks' Check Clearing Network
Average Percentage of Item Volume

	By Bank Assets in Millions of Dollars		
	Less than $500	**$500 to $4,999**	**$5,000 or More**
Local Checks			
Up to 1 business day	83.7	85.9	93.8
2 business days	12.7	11.0	5.9
More than 2 business days	3.5	3.1	0.3
Number of banks responding	159	61	29
Nonlocal Checks			
Up to 1 business day	42.2	53.2	65.7
2 business days	40.8	31.1	24.3
More than 2 business days	17.0	15.7	10.0
Number of banks responding	159	60	26

Source: American Bankers Association (1994).

Federal Reserve regulations. In their current form, these legal restrictions require that checks presented before the relevant cutoff time be paid at par on the same day.[2] The result is that paying banks do not compensate collecting banks for the interest lost while a check clears. Legal restrictions effectively mandate that John's bank is rewarded with an extra day's interest if it clears a check one day faster. The check float puzzle is thus an artifact of legal restrictions that mandate par presentment.

A deeper puzzle remains, however. Can we identify any economic benefits to Helen and her bank from faster check clearing? Are they large enough to warrant the interest earnings captured by presenting faster? The answer, as I will argue below, appears to be *no*.

Note that it is irrelevant how Helen and her bank divide between them the additional interest earnings due to check float. The question is why Helen and her bank, taken together, would want to compensate John and his bank (or someone presenting the check on their behalf) for presenting the check early. Similarly, it is irrelevant how John and his bank divide between them

[2] Under Regulation CC, checks presented by a depository institution before 8:00 a.m. on a business day must either be paid in reserve account balances by the close of Fedwire (currently 6:00 p.m.) or returned (12 CFR 229.36(f)). Under Regulation J, checks presented by a Reserve Bank before 2:00 p.m. on a business day must be settled the same day—the exact time is determined currently by each Reserve Bank's operating circular (12 CFR 210.9(a)).

the opportunity cost of foregone interest earnings. Taken together, they have an incentive to accelerate the presentment of Helen's check.

Some Efficiency Implications of the Allocation of Check Float

The check float puzzle would be merely an intellectual curiosity if it had little or no consequences for real resource allocations. Unfortunately, it appears that the allocation of check float earnings has a substantial effect on real resource allocation.

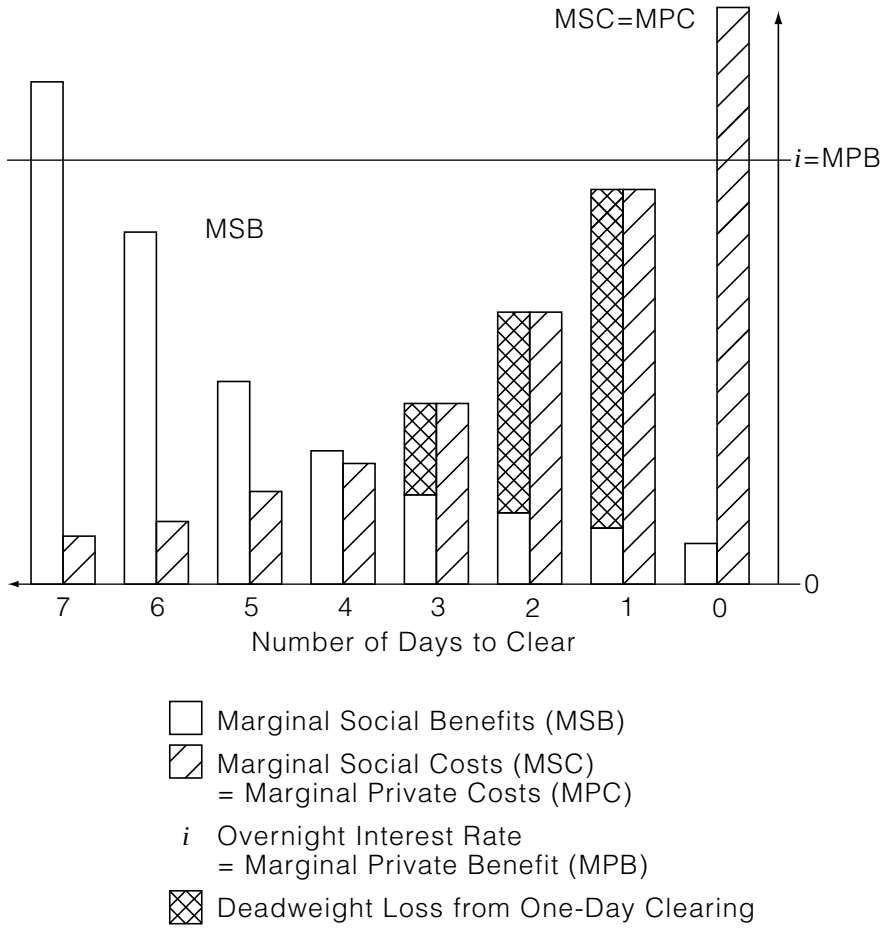
Consider the situation of John's bank, which has a range of options for clearing Helen's check. Some of these options are likely to differ in the speed with which they get the check to Helen's bank. Some clearing mechanisms might present the check in one day and some, particularly if Helen's bank is located far away, might take two or three days to present. The one-day methods have a distinct advantage for John's bank, because investable funds are obtained one day earlier. At the margin, John's bank is willing to incur real resource costs, in an amount up to one day's worth of interest earnings, in order to clear a check one day faster.

If, as I argue below, there is no identifiable social benefit of clearing a check one day faster, then the incremental resources expended to accelerate check collection and capture the interest earnings are wasted from society's point of view. The situation is illustrated in Figure 1. Check clearing speed is measured in days along the horizontal axis in Figure 1 and is increasing to the right. The position labeled "0" represents checks cleared the day they are first received, the position labeled "1" represents checks cleared one day after they are received, and so on. For a hypothetical check, the bars labeled *MPC* represent the marginal cost to the payees of clearing a check one day faster; the height for a clearing time of one day is the incremental cost of clearing in one day rather than two, the height for a clearing time of two days is the incremental cost of clearing in two days rather than three, and so on. Since these are real resource costs, they coincide with marginal social costs, so $MPC = MSC$. The marginal benefit to payees is measured by the horizontal line *MPB*; the height is the extra interest gained from earlier presentment.[3] If *MPB* exceeds *MPC*, the check is not being cleared too fast, from the payees' point of view, while if *MPC* exceeds *MPB*, the check is being cleared too fast. Payees will choose the fastest method of clearing checks that results in marginal benefits exceeding marginal costs.[4] For the checks portrayed in Figure 1, payees will

[3] I abstract from weekends, for which the extra interest would be three times as large as for weekdays.

[4] If interest compounds continuously and costs vary continuously with speed, then the payee bank would choose a method for which the marginal cost of accelerating presentment equaled the interest rate (*MB*).

Figure 1



present in one day; the marginal private cost of accelerating presentment in order to clear the same day exceeds the marginal private benefit.

I provide evidence below suggesting that the marginal social benefit of accelerating presentment is actually very small. Figure 1 therefore portrays the marginal social benefit curve *MSB* as relatively low for one-day clearing. Although the quantities in Figure 1 are not based on explicit empirical estimates, they are selected to illustrate the likely relative magnitudes involved. The socially optimal speed of check clearing in Figure 1 is four days; clearing any faster incurs marginal social costs that are greater than marginal social

benefits. The gaps between *MSC* and *MSB* between four days and one day—the cross-hatched bars—represent the deadweight social loss associated with the way check float earnings currently are allocated, as compared to a hypothetical arrangement that results in the optimal clearing time. In this sense the deadweight loss is "caused" by our existing check float arrangements.

The value of daily check float provides an upper bound on the incentive to expend resources to accelerate presentment. A rough calculation gives a sense of the potential magnitudes involved. The total value of the checks cleared in 1995 was approximately $73.5 trillion, or an average of $201 billion per day (Committee on Payment and Settlement Systems of the central banks of the Group of Ten countries 1995). The overnight interbank interest rate averaged 5.83 percent that year, which corresponds to 0.016 percent per day. Multiplying this overnight rate by the value of checks cleared yields $32.2 million per day ($201 billion times 0.000160), or $11.7 billion per year. This works out to about $0.18 per check, and represents the amount of real resource costs that would willingly be incurred by payees, like John and his bank, to present their checks one day faster. This corresponds to the height of the marginal private benefit line (*MPB*) in Figure 1. Since payee banks will ensure that *MSC* does not exceed *MPB*, it follows that *MSC* could be as large as $0.18 for the average size check. If, as I argue below, *MSB* is close to zero, then the cross-hatched bar for day 1 in Figure 1 is likely to be close to $0.18, or $11.7 billion in total. For comparison, Kirstin Wells (1996) estimates that the total cost to banks of processing and handling checks is between $0.15 and $0.43 per item.[5] If the marginal social benefits of accelerating presentment by a day are close to zero, then a substantial proportion of bank and payee processing costs could represent socially wasteful expenditures. Moreover, additional resources might be saved by clearing checks in three or more days, as illustrated in Figure 1 by the cross-hatched bars, for a time to presentment of two and three days.[6]

The prices of private package delivery services—United Parcel Service (UPS) and Federal Express—provide another rough guide to the cost of accelerating check presentment. The major services offer different delivery speeds at different prices. Assuming that prices in these relatively competitive businesses closely reflect costs, the price of overnight delivery can be compared to the price of slower delivery options to provide a crude estimate of the relative

[5] These estimates are only an upper bound on the relevant cost figures since they include the processing costs associated with receiving checks at paying banks.

[6] Note that float earnings (*MPB*) vary in proportion to the face value of the check, while costs generally do not. Marginal social benefits from reduced fraud losses are probably at least proportional to the face value of the check. Thus if payees are able to choose different clearing methods for different checks, then for large value checks the *MPB* and the *MSB* curves will be shifted upward, while the *MPC* curve will stay fixed. If it is too costly for payees to discriminate between checks, it is the average values of *MPB* and *MSB* that are relevant.

cost of overnight presentment and slower presentment.[7] The analogy between check presentment and package delivery is certainly imperfect; check presentment deadlines do not precisely match package company delivery deadlines, the items being shipped have different physical properties, and the package companies are able to track shipments in real time. Nonetheless, there are important similarities that make the comparison useful. Both use the same transportation technologies—airplanes and trucks. Both involve substantial sorting en route. And both process substantial volumes—63 billion checks annually (bundled together in packages) versus over 900 million items annually for Federal Express and 180 million items annually for UPS. In fact, both UPS and Federal Express contract with check processing firms to transport and present checks for them.

Table 2 displays sample shipping costs for UPS and Federal Express from Richmond, Virginia, to various locations. The Federal Reserve presents checks to all these locations by 2:00 p.m. the next day at the latest. For UPS letter delivery, delaying delivery by 25 $\frac{1}{2}$ hours, from 10:30 a.m. the next day to noon the second day, saves over 30 percent of the cost of next-day delivery. Delaying next-day delivery until late the second day (yielding third-day funds availability under current check presentment rules) saves about half the cost, while delaying delivery until late the third day (fourth-day funds availability) saves about 60 percent of the cost. For a one-pound package with UPS, delaying delivery to the third day saves about 70 percent of the costs. For a one-pound package sent via Federal Express, the savings are even larger. Delivery late the second day (third-day funds availability) reduces costs by almost 80 percent. These figures suggest that delaying check presentment could eliminate a substantial portion of check processing and handling costs.

Rough empirical calculations indicate, therefore, that current check float arrangements impose potentially significant social costs on the payment system. Are there offsetting social benefits?

Some Attempts to Explain the Check Float Puzzle

Eliminating Nonpar Presentment

As mentioned above, the presentment of checks is governed by legal restrictions that require that checks be paid at par on the day they are presented (see

[7] The analogy assumes that the price of delivery within a certain time frame closely approximates the average cost of delivery within that time frame. One potential weakness of this analogy is the possibility that there is a large fixed cost component and that the price differentials reflect different demand elasticities rather than different average costs. Price differentials are nonetheless limited by incremental and stand-alone costs; for either delivery option, slow or fast, the price must lie above the incremental cost and below the stand-alone cost for prices to be efficient and sustainable: see Weinberg (1994). If the demand for fast delivery is less elastic, as one might expect, then the price for slow delivery will lie close to the incremental cost of slow delivery, in which case the price differential will be no less than the difference in incremental costs.

Table 2 Shipping Rates from Richmond, Virginia
in dollars

Destination	Next day 10:30 a.m.	Second day noon	Second day close of business	Third day close of business
UPS: Letter				
Baltimore	$11.00	$ 7.50	$ 5.75	$ 4.40
Birmingham	12.50	8.00	6.25	4.90
San Francisco	13.50	9.50	7.25	5.80
UPS: One-Pound Package				
Baltimore	$14.00	$ 7.75	$ 6.25	$ 4.40
Birmingham	17.25	8.25	6.75	4.90
San Francisco	20.00	10.50	8.25	5.80

Federal Express: One-Pound Package (all locations)		
Next day 8:00 a.m.	**Second day 10:30 a.m.**	**Second day 4:30 p.m.**
$47.50	$22.50	$ 9.95

Sources: United Parcel Service (1997); Federal Express Corporation (1996).

footnote 2). Do such legal restrictions serve any efficiency-enhancing role that might justify the inefficiencies caused by excessively rapid check presentment?

The current system of presentment regulations arose over the last 90 years since the founding of the Fed. Before the Fed was established in 1914, many banks charged presentment or "exchange" fees on checks sent to them for payment. Some state laws at the time held that a check presented "over the counter" shall be paid at par, but presentment fees could be charged when the collecting bank presented by indirect means, such as by mail. The banks charging presentment fees (so-called nonpar banks) were often small and rural, and they justified their fees as a way of covering the cost of remitting funds by shipping bank notes to the collecting bank.[8]

In drafting the Federal Reserve Act, the Reserve Banks were given the power to clear and collect checks, in part to help attract members to the Federal Reserve System (Stevens 1996). While national banks were required to become members, few state-chartered banks joined the System in the early years. At

[8] The term par presentment is generally taken to refer broadly to the right to present by indirect means such as mail or courier service and still receive par.

first the Reserve Banks tried a voluntary clearing system in which they accepted at par only checks drawn on other members who agreed to accept checks at par. This scheme failed to attract enough participants and was abandoned after a year in favor of the somewhat misnamed "compulsory" system in July 1916.[9] Under the new scheme Reserve Banks accepted checks drawn on any member banks or on nonmember banks that agreed to accept checks at par. The Reserve Banks campaigned hard to get banks to agree to accept at par and had greater success. Congress helped by revising the Federal Reserve Act in 1917, adding a provision that no presentment fees could be charged against the Fed, although specifically authorizing "reasonable charges" against other presenting banks. The Reserve Banks thus acquired the unique legal privilege of being able to present at par by indirect means, such as by mail. Membership increased dramatically in the years that followed, and the Reserve Banks were successful in significantly curtailing, though not eliminating, nonpar banking. Presentment fees were effectively eliminated in 1994 when the Fed introduced regulations that mandated same-day settlement for checks presented by 8:00 a.m.

The conventional view is that par presentment regulations were instrumental in allowing the Fed to enter the check clearing business and that this enhanced the efficiency of the check collection system. If so, then eliminating inefficiencies in check collection represents a social benefit that might outweigh the social waste due to excessively fast presentment. One potential explanation of the check float puzzle, then, is that it reflects a side effect of a par presentment regime whose net social benefits are positive.

Two types of claims have been made about the efficiency-enhancing role of par presentment. The first argument, advanced by contemporary observers just after the founding of the Fed, was that presentment fees resulted in wasteful practices on the part of collecting banks seeking to avoid them. After the check is written and accepted in payment, the paying bank has a monopoly on the ability to redeem the check. Paying banks would set charges well above costs to extract rents from collecting banks (Spahr 1926). Payee banks would in turn try to avoid paying what they saw as exorbitant fees. A bank typically would have a network of correspondent banks with whom it exchanged checks. A correspondent bank would present checks directly on behalf of the sending bank or would send the check on to another correspondent, hoping it had an arrangement for direct presentment. The second correspondent might then send the check further on, and so forth. Checks sometimes traveled circuitous routes as banks sought a correspondent whom they hoped would allow them to avoid presentment charges (Cannon 1901). Such practices, it was asserted, resulted in wasteful shipping costs and inefficient delay in payment.

[9] One reason the voluntary scheme failed was the policy of crediting and debiting banks immediately when checks were received. There was a lag before banks were informed of debits, which made reserve management difficult and overdrafts frequent.

A second argument for the efficiency-enhancing role of par presentment is advanced by modern critics of the pre-Fed check collection system. Unilaterally set presentment fees allow a bank to increase retail market share by raising the costs of rival depository institutions (McAndrews and Roberds 1997; McAndrews 1995). Nonpar banking allows a "vertical price squeeze" in which a bank inefficiently raises the price of an upstream input (presentment) purchased by a bank that is a rival in a downstream market (retail deposit-taking).[10] Presentment fees are an anticompetitive practice, according to this argument, and the establishment of par presentment eliminated the associated inefficiencies.[11]

These two arguments fail to explain the check float puzzle. Regarding the first argument, it is not at all obvious that nonpar banking was inefficient. It is important to note that a collecting bank was not completely at the mercy of the paying bank. Collecting banks always had the option of finding a correspondent to present directly on their behalf, thereby avoiding the presentment fee. Competition between correspondent banks ultimately governed the cost of clearing checks drawn on distant banks and placed a ceiling on the presentment fees banks could charge. Moreover, the occasional circuitous routing of checks is not obviously inefficient, given the necessity of relying on a network of bilateral relationships (Weinberg 1997). It is a common feature of network transportation and communication arrangements; after all, the circuitous routing of telephone calls is not taken as evidence of inefficiency.

Another common feature of network arrangements is the presence of fixed costs. In such settings there typically is a range of prices consistent with efficiency and sustainability. Each participant obviously will prefer to bear as little of the fixed costs as possible. Critics of presentment fees wanted paying banks to bear more of the common costs of check clearing. Defenders of presentment fees wanted collecting banks to bear more of the costs. The par presentment controversy appears to have had more to do with distributional issues than with economic efficiency.

The view that presentment fees can facilitate a vertical price squeeze is based on models that take many important aspects of the institutional arrangements governing check clearing as fixed. Models in which such arrangements are endogenous can have very different predictions. For example, Weinberg (1997) describes a model of check clearing in which outcomes are efficient, even without restrictions on presentment fees. Such models are attractive in this setting because, historically, check clearing has often involved cooperative arrangements between banks, such as clearinghouses. Moreover, the banks most susceptible to a vertical price squeeze by the nonpar banks were located close

[10] See Salop and Scheffman (1983) for a basic exposition, and Laffont (1996) and Economides, Lopomo, and Woroch (1996) for applications to network industries.

[11] McAndrews (1995) argues that the imposition of any uniform presentment fee would suffice to eliminate this inefficiency.

by, and were the very banks that could present directly. The banks that bore the brunt of presentment fees were those located at a distance and thus least likely to lose retail customers to the paying bank.

More to the point, check clearing arrangements provided the same incentives to accelerate presentment both before and after the founding of the Fed. Under state laws and established common law principles, the presenting bank was entitled to immediate payment at par for checks presented over the counter. Thus a bank presenting directly to the paying bank faced the same relative incentives before and after the entry of the Fed into check clearing; getting the check there one day earlier resulted in one day's worth of interest. Over-the-counter presentment served as an anchor for the prices of other means of presentment. It placed a bound on the payee bank's willingness to pay an exchange fee for presenting by mail or to pay a correspondent bank for collecting the check. Neither the paying bank nor the correspondent bank had any incentive to compensate the payee bank for the interest foregone before remitting the check. Thus the relevant property of the par presentment regime predates the Fed's entry into check clearing. The elimination of nonpar presentment cannot explain the check float puzzle.

Reducing Check Fraud

Another possible explanation of the check float puzzle is that clearing checks faster reduces check fraud losses to paying banks and their customers. Helen's bank might be willing to compensate John's bank for getting the checks to them sooner because it reduces the expense associated with check fraud.

There are various ways in which banks and their customers can lose money to check fraud. Someone possessing lost or stolen checks can forge the account holder's signature or the endorsement. Checks can be altered without the account holder's approval. Counterfeit checks resemble genuine checks and can sometimes be used to obtain funds. Checks can be written on closed accounts. Fraudulent balances can be created through "kiting"—writing a check before covering funds have been deposited.

When Helen's check is presented for payment her bank can verify the signature and the authenticity of the check and can verify that the account contains sufficient funds. If her bank chooses to dishonor the check, it must initiate return of the check by midnight of the business day following the day the check was presented. The check is then returned to John's bank. If Helen's bank paid the check when it was presented, then a payment is made in the opposite direction when the check is returned. Otherwise Helen's bank returns the check without paying.

Note, however, that if Helen's bank returns the check, Helen's bank bears no loss. John and his bank now have a check that was dishonored, and between them they bear the loss (or else seek compensation from Helen). John and his

bank can be expected to take into account the effect of the speed of check clearing on the likelihood of their fraud losses. Therefore, the losses experienced by payees and their banks do *not* help explain the check float puzzle. The losses that *are* relevant to our puzzle are those borne by Helen and her bank. They would be willing to compensate John's bank to induce more rapid clearing if that helped reduce their own check fraud losses.[12]

There are a number of reasons why check fraud losses to the paying bank might be reduced if it received the check faster. Helen's bank may allow the time limit for check returns to elapse before finding out that the check is forged or that Helen has closed her account. Some banks, for example, do not routinely verify signatures. In this case, Helen's bank bears the loss. Such losses might be lower for checks presented faster. Helen's bank might want to provide an implicit reward to John's bank for rapid presentment. In principle, then, the desire to encourage rapid check clearing to discourage check fraud might explain the check float puzzle.

But is the check fraud effect large enough empirically to explain the check float puzzle? Does getting the check to Helen's bank one day faster reduce fraud losses at Helen's bank by enough to justify providing John's bank with one more day's interest on the funds? According to a recent Board of Governors report to Congress (Board of Governors 1996), check fraud losses incurred by U.S. commercial banks, thrifts, and credit unions amounted to $615.4 million in 1995. Some check fraud losses occur to banks in their role as collectors of checks drawn on other banks, and some occur to banks in their role as payors of checks drawn on other banks. Of the total estimated check fraud loss mentioned above, only about half—$310.6 million—represents losses to banks as payors. The remainder represents losses to banks as collectors. As noted above, only check fraud losses to the payor are directly relevant to the check float puzzle.

The figures just cited are *gross* losses, however. The Board study reports that depository institutions *recovered* a total of $256.0 million on past check fraud losses in 1995, although it does not indicate how these recoveries were divided between paying banks and collecting banks. If we take these as estimates of steady-state losses and recoveries, and if we assume that recoveries are the same fraction of gross losses for both collecting banks and paying banks, then paying banks experienced net check fraud losses of $181.4 million in 1995.[13] Average net check fraud losses at paying banks therefore amounted to less

[12] Figure 1 could be modified to account for the desire of John and his bank to reduce their check fraud losses. The marginal benefit from reducing their expected losses should be added to the marginal private benefit curve *MPB*. The same amount should be added to the marginal social benefit curve, *MSB*, as well, so the net distortion remains the same.

[13] Recoveries by paying banks are (50.5%) $\times$ ($256.0 million) or $129.2 million, so net losses are $310.6 million minus $129.2 million, or $181.4 million. Note that the resulting figure is conservative in the sense that if check volume is growing, then this procedure underestimates the ratio of recoveries to gross losses.

than 0.0003 cents per dollar in 1995.[14] In comparison, one day's interest on the check, at a 5.5 percent annual rate (the current overnight Fed funds rate), is worth 0.015 cents per dollar; more than 50 times as large as the average rate of net check fraud losses at paying banks.

The check fraud loss figure is the *average* net loss, however. The relevant figure is the *marginal* effect on net fraud loss of clearing a check one day faster. It could conceivably be the case that, say, the expected fraud loss on a check cleared in two days exceeds the expected loss on a check cleared in one day by 0.015 cents per dollar, the value of the float, even while the average check fraud loss is 0.0003 cents per dollar. Unfortunately, there are no figures available that would allow us to estimate directly marginal net fraud losses. However, for the average net expected loss to be as small as 0.0003 cents while the marginal loss associated with clearing a check in two days rather than one day is as large as 0.015 would require that no more than 2 percent of checks take two or more days to clear.[15] No more than 2 percent is quite implausible, however, given the figures in Table 1, which show that a substantial portion of checks take two days or more to clear. Thus, even though we do not have a direct measure of the marginal expected fraud loss associated with clearing a check one day slower, the evidence strongly suggests that fraud loss at paying banks does not explain the distribution of check float earnings.

Check writers themselves sometimes suffer losses due to check fraud. Perhaps Helen's desire to limit her own check fraud losses makes her and her bank willing to forego the extra interest earnings in order to induce more rapid clearing of her checks. There are two principal methods by which a depositor could lose money due to check fraud. One is if Helen fails to inspect periodic bank statements for forged or unauthorized checks, she can be apportioned

[14] Calculated as $181.4 million divided by $73.5 trillion (dollar value of checks written in 1995 [Committee on Payment and Settlement Systems of the central banks of the Group of Ten countries 1995]) = 0.0003.

[15] Let α_i be the fraction of checks (by value) cleared in i days, and let γ_i be the expected fraud loss on checks cleared in i days. Expected fraud loss is then $\alpha_1\gamma_1 + \alpha_2\gamma_2 + \ldots = 0.0003$. Suppose, hypothetically, that the marginal loss associated with clearing one extra day, $\gamma_{i+1} - \gamma_i$, is at least 0.015. What values of α_1 are consistent with these two assumptions? The most optimistic case, in the sense that the allowable range for α_1 is the largest, is one in which all checks clear in either one or two days, because the longer it takes to clear the larger the expected loss. As long as $\gamma_{i+1} \geq \gamma_i$, the best case is for α_i to be as small as possible for $i \geq 3$, because increasing the weights on the days with larger losses makes it harder to match the average loss figure of 0.0003. Assume therefore that $\alpha_i = 0$ for $i \geq 3$. Similarly, the most optimistic assumption to make about γ_1 is $\gamma_1 = 0$, because increasing γ_1, the expected loss on the smallest loss day, just makes it harder to match the average loss figure. Our two postulates are now $(1 - \alpha_1)\gamma_2 = 0.0003$, and $\gamma_2 \geq 0.015$, which together imply that $1 - \alpha_1 \leq (0.0003/0.015) = 0.02$.

Looked at another way, for given fractions α_i, how large can $\gamma_2 - \gamma_1$ be and still satisfy $\alpha_1\gamma_1 + \alpha_2\gamma_2 + \ldots = 0.0003$ and $\gamma_{i+1} \geq \gamma_i$? The answer is $0.0003/(1 - \alpha_1)$. From the figures in Table 1 this ranges from 0.0005 to 0.005, or 3.5 to 32.3 percent of the monetary value of one day's worth of float.

some of the loss on grounds of negligence. But the timeliness of check clearing is only marginally important in such cases, since they involve inspecting monthly bank statements.

Another method by which a depositor could lose money involves "demand drafts," one-time pre-authorized checks written by merchants or vendors after taking a depositor's bank account number over the phone. In place of the customer's signature the check is stamped "pre-approved" or "signature on file." Demand drafts are cleared the same way as conventional checks and have many legitimate uses, but they have been used in telemarketing scams. It seems unlikely that the detection and prosecution of such fraud depends significantly on the speed with which demand drafts are cleared. Most cases seem to be discovered when a depositor's bank statement is inspected. Moreover, such fraud only affects demand drafts, and these are a tiny fraction of all checks written.[16] So in neither case does fraud loss by check writers appear to be a plausible rationale for the allocation of check float earnings.

There is an additional reason to doubt that fraud losses could ever explain why the collecting bank should lose interest earnings until the check is presented. The relevant interest rate is the nominal overnight rate, and thus will vary directly with expected inflation, other things being equal. There is no reason why the additional expected fraud loss associated with clearing a check in two days rather than one should have any necessary relationship with the inflation rate. Indeed, the inefficiency caused by the fact that checks do not bear interest parallels exactly the traditional welfare cost of anticipated inflation, which is caused by the fact that currency does not bear interest. The inefficiency of currency use arises because people go to excessive lengths to avoid holding it. Similarly, check float arrangements cause banks to go to excessive lengths to avoid holding checks. In both cases the problem is that the rate of return is artificially depressed by inflation. The difference between the two is that, apart from changing the inflation rate, altering the rate of return on currency, say by paying interest, appears to be technologically difficult. In contrast, as I argue below, the technology to alter the rate of return on checks appears to be readily available.[17]

The Expedited Funds Availability Act

When an account holder deposits a check at a bank, the common banking practice is to place a "hold" on the funds for a number of days until the bank is

[16] Legitimate demand drafts probably amount to less than $1 billion a year. Jodie Bernstein, Director of the Bureau of Consumer Protection, reported one estimate that "nine of the current twenty demand draft service bureaus process approximately 38,000 demand drafts weekly, totaling over five million dollars. . . ." In other words, $250 million annually (Bernstein 1996).

[17] Reducing inflation to the socially optimal rate would accomplish the desired objective, but I take that as outside the realm of check regulatory policy.

certain that the check has cleared. The bank customer is not allowed to withdraw the funds until the hold is removed. This practice protects the bank from fraud by shifting some of the risk to the account holder. In 1987 Congress passed the Expedited Funds Availability Act (EFAA), which asked the Federal Reserve to promulgate regulations limiting the length of time banks can hold customers' funds. Maximum holds vary from one to five business days, depending on the type of check and whether or not it is a "local" item.

Legal restrictions on the duration of holds can be an incentive to accelerate check presentment. After the hold is released, the funds may be withdrawn, and the bank may suffer a loss if the check is returned unpaid. Does this explain the check float puzzle? The answer is clearly no. Congress enacted the EFAA to respond to concerns that holds were longer than were necessary to ascertain whether the check would be returned unpaid. The EFAA explicitly instructs the Federal Reserve Board to reduce the allowable time periods to the minimum consistent with allowing a bank to "reasonably expect to learn of the nonpayment of most items." The hold periods, in other words, are tailored to the speed with which checks are actually being collected, not the other way around.

The EFAA constrains the distribution of the risk of nonpayment between the payee and the payee's bank. But it does nothing to alter the incentive both parties have to take steps to reduce their joint losses from fraud. The EFAA does increase the ability of payees to perpetrate fraud on their banks and so provides an extra incentive for payee banks to accelerate presentment. If the EFAA artificially discouraged faster presentment, such discouragement might explain the need for the compensating stimulus provided by the current check float arrangement. But if anything, the EFAA heightens the incentive to accelerate presentment.

What Can Be Done?

I conclude that the social benefit of accelerating check presentment is negligible in comparison to the reward to collecting banks in the form of captured interest earnings. Apparently this feature of the check clearing system does not have an identifiable economic rationale. Without any offsetting social benefits, we are left with just the social costs described earlier.

Is there an alternative to the current arrangements governing check float? Is there a practical way to eliminate the artificial incentive to accelerate the presentment of checks? After all, it could be the case that the current scheme has deadweight social costs but is superior to all feasible alternatives. Is there a feasible alternative that does not require the deadweight social costs noted above?

Consider first what properties an ideal arrangement would possess. In an ideal arrangement the value to John's bank of presenting a check one day sooner would equal the real value to Helen and Helen's bank of receiving the check one day sooner. Fraud losses (to the payor bank) aside, John's bank should

implicitly earn interest on the check while it is being cleared. Helen's bank should implicitly pay interest to John's bank from the time at which John's bank received the check. John's bank would then face no artificial inducement to accelerate presentment. Note that John's bank still has an incentive to clear the check, since fraud losses to the payee bank are likely to increase the longer it takes to clear the check. But the magnitude of the incentive to accelerate presentment would match the social value of accelerating presentment.

Check fraud losses to the payor bank constitute an additional social value of accelerating presentment. To account for these precisely, the implicit interest rate on checks should be reduced by the marginal effect of delaying presentment on payor fraud losses, resulting in a slight penalty for delaying presentment. As noted previously, however, the marginal effect on payor bank fraud losses is likely to be quite small when compared to the interest earnings at stake. In an ideal arrangement, therefore, we should see checks in the process of collection implicitly bearing interest at close to the overnight rate.

Implementing an ideal arrangement would require revising the current par presentment regulations. One possibility is to have the paying bank pay explicit interest on the face value of the check from the date the check was originally accepted by the bank of first deposit. The interest would be paid directly to the presenting institution. The interest rate could be determined by reference to a publicly available overnight rate. Regulations would stipulate that upon presentment, the paying bank is accountable for the amount of the check plus accrued interest from the date of first deposit. The regulation would constrain only the obligations of the paying bank. If the collecting bank was presenting on behalf of some other bank, they could divide the interest between them as they see fit. Presumably each bank would receive the interest accruing while the check was in their possession. Similarly, the regulation would be silent on the division of interest between the bank of first deposit and its customer.

A second possibility is for checks to be payable at par only at a fixed maturity date—say, five business days after the check is first deposited in a bank. Checks presented before five business days would be discounted, again using a publicly available overnight interest rate as reference. After five days an outstanding check would accrue interest at the reference rate. The maturity date would determine the implicit division of revenues between paying banks and payee banks.

The main practical difficulty facing any such scheme is to record and transmit the date on which the check is first deposited. Currently, the Federal Reserve's Regulation CC requires that the bank at which the check is first deposited print on the back of the check certain information (the indorsement), including the date. This information is used mostly in the process of returning checks and is not machine-readable. Some information on a check is machine-readable, however. At some point early in the clearing process, the dollar amount is printed in magnetic ink on the bottom of the check front beside

the paying bank's routing number and the payor's bank account number. The resulting string of digits and symbols—the so-called "MICR line" at the bottom of the check—is read automatically as the check subsequently is processed. One possibility would be to expand the MICR coding format to include the date as well. Then the implicit interest obligation could be handled using the same automated techniques used to handle the face amount. Although this alternative regime would certainly involve transitional costs, the figures discussed above indicate that the potential benefits are substantial—perhaps as large as billions of dollars per year.

Note that this proposal would have the side benefit of facilitating improved contractual arrangements between banks and their customers by giving them more readily usable information on when a check was cleared. This information could be used by banks to penalize kiting if they so desired. Banks might charge check writers for the interest paid to the bank presenting a check. The arrangement would be a matter of contractual choice for banks and their customers, however, and would not affect the desirability of the proposal.

In the Meantime, There Are Some Important Implications

Until we establish a more rational scheme for allocating check float earnings, payment system policymakers apparently face a dilemma. They are often asked to contemplate changes to the payment system that would alter the speed with which some checks are cleared. One example is a proposal to close down the Fed's Remote Check Processing Centers (Benston and Humphrey 1997). This would likely slow down the collection of some checks. Another example is a proposal for electronic check presentment (ECP), which involves transmitting electronically to paying banks the encoded information on checks (Stavins 1997). In this case, checks would likely be collected somewhat faster on average.

How should such changes in check float affect the decision? One point of view (the "zero-sum view") asserts that the change in float earnings is merely a transfer. The gain realized by payees and their banks from faster presentment is exactly matched by a corresponding loss to payors and their banks. In this view, changes in float should be ignored in policy analysis. That is, in a social cost-benefit analysis, no weight should be given to changes in float. This view is in accord with the evidence cited above that the social benefit of accelerating check clearing is negligible.

The danger in this approach, however, is that payment system participants respond to the (distorted) incentives embodied in the current arrangements; consequently their reactions could be misgauged. Imagine that the Fed is considering a change that would increase check float. For example, suppose that the closure of an RCPC slowed down the collection of some deposited checks. For the checks the Fed continues to process, the slowdown would reduce the

amount of resources wasted on accelerating presentment. But it would do nothing to reduce the incentive banks have to accelerate presentment. Banks could respond by clearing directly themselves or through private service providers, rather than through the Fed, in order to minimize float. If the social cost of clearing checks outside the Fed is greater than the cost of clearing them through the Fed, then there might be no net social savings to closing down the RCPC, since the increase in private costs might outweigh the decrease in Fed costs. A cost-benefit analysis that ignored the effect of changes in float could be seriously misleading.

An alternative approach (the "empirical view") would treat the overnight interest rate as the social value of accelerating presentment, as if there is some as-yet-undiscovered social benefit of reducing check float. This approach has the advantage of aligning policy objectives with the incentives faced by private participants in the check collection industry. The danger in this approach is the risk of favoring speedy check presentment when it is not really in society's best interest. Suppose again that the Fed is considering closing an RCPC, but that no banks switch to other means of clearing checks. The increase in float would be counted against closing the facility, under the empirical view. It could turn out that, if one disregards the increased float, then the net social benefits of closing the facility are positive (due to the resources saved by clearing more slowly) but are negative when the value of the lost interest earnings to payee banks is deducted.[18] In this case, the empirical approach recommends against closing the facility even though it really should be closed. By adopting the empirical view, policymakers would be joining in the private sector's wasteful pursuit of float.

The dilemma is more apparent than real, however. Policymakers should focus on the implications for real resource costs of the proposals they are considering and should exclude the purely pecuniary impact of reallocations of check float. But they should keep in mind that although float does not reflect any direct social benefits, it does affect behavior. To the extent that reallocations of float induce behavioral changes that alter real resource use, the induced changes in resource costs must be included in any cost-benefit analysis.

Current float arrangements can be thought of as imposing a tax paid by presenting banks on checks cleared by slower methods, with the proceeds automatically passed on to payor banks. The proper treatment of a tax in cost-benefit analysis is well understood. Absent other interventions, the taxed service (slow clearing) will be undersupplied relative to the untaxed service (fast clearing) for which it is a substitute. If a public entity like the Fed is active in supplying the untaxed good, and unilaterally cuts back on its supply, providing more of

[18] The float that Reserve Banks experience is passed back to depositing banks. If, for example, 97 percent of a particular class of checks is cleared in one day and the rest in two days, on average, depositors receive 97 percent of their funds in one day and the rest in two days.

the taxed good instead, the net effect will depend on the market for the untaxed good. At one extreme, the Fed might have many competitors whose costs and prices are close to that of the Fed. In this case reducing the supply of the untaxed service merely causes customers to switch to competitors—no improvement in efficiency results. At the other extreme, if the Fed has few competitors for the supply of the untaxed service—no other suppliers have costs close to the Fed's—then customers can be induced to switch to the socially superior taxed good. Here, slowing down Fed check collection does not drive customers away, with the result that check collection does indeed slow down and thus saves societal resources. Note that this outcome could increase costs to Fed customers in the sense that Fed fees plus float costs increase, even though social costs decrease.

In the decision to close an RCPC, for example, the analysis should take into account the effect of increased float on depositing banks' check clearing choices. To the extent that increased float causes banks to switch to other providers—private check clearing services or correspondent banks, for example—the increase in the real resource costs of alternative check clearing operations should be counted against any savings in real resource costs associated with Fed check clearing. The change in float earnings itself should be excluded from the calculation of net social benefits, but the effect on bank choices must be taken into account.

In evaluating ECP, the float benefits to payees from faster presentment should not count as a social benefit, as Joanna Stavins (1997) correctly points out. If ECP is offered under current par presentment regulations, however, the benefits of float arising from faster presentment (assuming they are passed back to depositing banks, as is current Fed practice) would be an artificial stimulus to the adoption of ECP. If ECP is offered at prices that are efficient (relative to the real resource costs of ECP) and the extra float earnings from faster presentment are passed on to payees, then ECP may be adopted where it is not socially efficient.[19] For some checks ECP might be more costly than physical presentment, and yet customers would prefer ECP because of the benefits of reduced float. The Fed should avoid deploying ECP in market segments where it would increase social costs, even if it would decrease Fed customers' costs (including float costs).

More generally, the check float problem can distort the process of technological innovation by artificially promoting techniques that accelerate check presentment. Payment system participants have an incentive to find new ways

[19] ECP with check truncation is often said to involve "network effects" because such a scheme would be most valuable if universally adopted, eliminating all paper presentment. The same logic applies, however. The set of prices that are efficient and sustainable relative to resource costs alone will not in general coincide with the set of prices that are efficient relative to the aggregate of resource costs and float costs. See Weinberg (1997) regarding network effects in payment arrangements.

to reduce their holdings of non-interest-bearing assets, like currency and checks (Lacker 1996). This incentive is merely an artifact of the inflation tax, and thus does not represent any fundamental social benefit (Emmons 1996). The check float problem is another example of the way inflation can distort the payment system.

The check float puzzle has important implications for the role of the Federal Reserve in the check clearing industry. The Fed currently enjoys certain competitive advantages over private participants. One involves the disparity in presentment times mentioned above; the Fed can present until 2:00 p.m. for same-day funds, while others must present before 8:00 a.m. for same-day funds (unless varied by agreement). This disparity gives the Fed a competitive advantage, because depositors can be offered a later deposit deadline at a cost lower than that of a private provider. Having such a competitive advantage would allow the Fed, should it so desire, to improve the efficiency of check collection by slowing down presentment and increasing check float beyond that which the private market would provide.[20] It gives the Fed an ability to offset some of the deleterious side effects of par presentment regulations. Note that this outcome is the opposite of the original justification of the Fed's role in check clearing provided by opponents of presentment fees, who claimed that the Fed would result in more rapid check clearing.

The Fed's advantage over private providers of check clearing services has been eroding over time. In 1980 Congress passed the Monetary Control Act, which required that the Fed charge prices for its payment services comparable to those that would be charged by private providers. Effective in 1994, Regulation CC was amended to allow "same-day settlement"—private presentment as late as 8:00 a.m. for same-day funds. Because of these changes and other factors, the Fed's market share has been steadily eroding in recent years (Summers and Gilbert 1996). Payment system efficiency no doubt helped motivate this movement towards a "level playing field." And yet these changes have reduced

[20] To see this, consider the following simplified situation. The Fed faces private providers with costs of γ_1 of clearing a check in one day and γ_2 of clearing a check in two days. The value of one day's float on a typical item is i. Under competitive conditions the cost to a depositor is $\gamma_1 + i$ for clearing privately in one day, and $\gamma_2 + 2i$ for clearing privately in two days. Clearing in two days is socially optimal, so $\gamma_1 > \gamma_2$, there being no other relevant social costs or benefits associated with check clearing. But under the current regime checks are collected (inefficiently) in one day; that is, $\gamma_1 + i < \gamma_2 + 2i$, or $\gamma_1 - i < \gamma_2$. The Fed offers check clearing, but only two-day clearing. Suppose the Fed's cost of clearing in two days is δ_2, and the Fed charges p per item. Cost recovery requires (a) $p \geq \delta_2$. Can the Fed attract depositors that are now clearing privately in one day? This requires (b) $p + 2i < \gamma_1 + i$. Together, (a) and (b) are feasible if $\delta_2 < \gamma_1 - i < \gamma_2$. The Fed's presentment time advantage implies that the Fed can present checks in a given number of days at lower cost than the private sector can present checks in the same number of days: in other words, δ_2 is strictly less than γ_2, as required. Thus the Fed's presentment time advantage allows the Fed to reduce check clearing time from one day to two days in this example, improving the efficiency of the check collection.

the Fed's ability to unilaterally improve the efficiency of check collection by slowing down check presentment.

Now is a good time, therefore, to reexamine the Fed's role in the check collection industry and the payment system more broadly.[21] As noted earlier, the rationale for the Fed's original entry into check collection was to improve efficiency. But the par presentment regulations that once aided the Fed's entry are now clearly an impediment to efficiency. Can the Fed still play an efficiency-enhancing role in the presence of par presentment regulations? Can the Fed implement technological improvements to the payment system without removing inefficient par presentment regulations? These questions should be at the heart of any reexamination of the Fed's role in the payment system.

REFERENCES

American Bankers Association. "1994 ABA Check Fraud Survey." Washington: American Bankers Association, 1994.

Benston, George J., and David B. Humphrey. "The Case for Downsizing the Fed," *Banking Strategies* (1997), pp. 30–37.

Bernstein, Jodie. "Demand Draft Fraud." Prepared Statement before the House Banking Committee: Federal Trade Commission, 1996.

Board of Governors of the Federal Reserve System. "Report to the Congress on Funds Availability Schedules and Check Fraud at Depository Institutions." Washington: Board of Governors of the Federal Reserve System, 1996.

_____. "Guidelines Approved for New Check-Clearing System," *Federal Reserve Bulletin,* vol. 58 (February 1972), pp. 195–97.

_____. "Statement of Policy on Payments Mechanism," *Federal Reserve Bulletin,* vol. 57 (June 1971), pp. 546–47.

Cannon, James G. *Clearing-Houses: Their History, Methods and Administration.* London: Smith, Elder, & Co., 1901.

Committee on Payment and Settlement Systems of the central banks of the Group of Ten countries. "Statistics on Payment Systems in the Group of Ten Countries." Basle: Bank for International Settlements, 1995.

Economides, Nicholas, Giuseppe Lopomo, and Glenn Woroch. "Strategic Commitments and the Principle of Reciprocity in Interconnection Pricing." New York: Stern School of Business, NYU, 1996.

[21] In October 1996 Federal Reserve Chairman Alan Greenspan appointed a committee, headed by Board Vice Chair Alice M. Rivlin, to review the Fed's role in the payment system.

Emmons, William R. "Price Stability and the Efficiency of the Retail Payments System," Federal Reserve Bank of St. Louis *Review,* vol. 78 (September/October 1996), pp. 49–68.

Federal Express Corporation. *Fedex Quick Guide*. Memphis: Federal Express Corporation, 1996.

Humphrey, David B. "Checks Versus Electronic Payments: Costs, Barriers, and Future Use." Manuscript. Florida State University, 1996.

_____, and Allen N. Berger. "Market Failure and Resource Use: Economic Incentives to Use Different Payment Instruments," in David B. Humphrey, ed., *The U.S. Payment System: Efficiency, Risk and the Role of the Federal Reserve*. Boston: Kluwer, 1990.

Lacker, Jeffrey M. "Stored Value Cards: Costly Private Substitutes for Government Currency," Federal Reserve Bank of Richmond *Economic Quarterly,* vol. 82 (Summer 1996), pp. 1–25.

Laffont, Jean-Jacques, Patrick Rey, and Jean Tirole. "Network Competition: I. Overview and Nondiscriminatory Pricing." Manuscript. 1996.

McAndrews, James J. "Commentary," Federal Reserve Bank of St. Louis *Review,* vol. 77 (November/December 1995), pp. 55–59.

_____, and William Roberds. "A Model of Check Exchange." Manuscript. Philadelphia: Federal Reserve Bank of Philadelphia, 1997.

Salop, Steven C., and David T. Scheffman. "Raising Rivals' Costs," *American Economic Review,* vol. 73 (May 1983, Papers and Proceedings), pp. 267–71.

Spahr, Walter Earl. *The Clearing and Collection of Checks*. New York: Bankers Publishing Co., 1926.

Stavins, Joanna. "A Comparison of Social Costs and Benefits of Paper Check Presentment and ECP with Truncation," *New England Economic Review* (July/August 1997), pp. 27–44.

Stevens, Ed. "The Founders' Intentions: Sources of the Payments Services Franchise of the Federal Reserve Banks." Cleveland: Financial Services Working Paper Series, 1996.

Summers, Bruce J., and R. Alton Gilbert. "Clearing and Settlement of U.S. Dollar Payments: Back to the Future?" Federal Reserve Bank of St. Louis *Review,* vol. 78 (September/October 1996), pp. 3–27.

United Parcel Service. "Quick Cost Calculator." Available: http://www.ups.com [April 1997].

Veale, John M., and Robert W. Price. "Payment System Float and Float Management," in Bruce J. Summers, ed., *The Payment System: Design, Management, and Supervision*. Washington: International Monetary Fund, 1994.

Weinberg, John A. "The Organization of Private Payment Networks," Federal Reserve Bank of Richmond *Economic Quarterly,* vol. 83 (Spring 1997), pp. 25–43.

_____. "Selling Federal Reserve Payments Services: One Price Fits All?" Federal Reserve Bank of Richmond *Economic Quarterly,* vol. 80 (Fall 1994), pp. 1–23.

Wells, Kirstin E. "Are Checks Overused?" Federal Reserve Bank of Minneapolis *Quarterly Review,* vol. 20 (Fall 1996), pp. 2–12.

A Review of the Recent Behavior of M2 Demand

Yash P. Mehra

It is now known that the public's M2 demand experienced a leftward shift in the early 1990s. Since about 1990 M2 growth has been weak relative to what is predicted by standard money demand regressions. It is widely believed that this shift in money demand reflected the public's desire to redirect savings flows from bank deposits to long-term financial assets including bond and stock mutual funds. Recognizing this, policymakers have not paid much attention to M2 in the short-run formulation of monetary policy since July of 1993.[1]

In this article, I review the recent behavior of M2 demand. I then evaluate the hypothesis that the recent shift in M2 demand can be explained if we allow for the effect of the long-term interest rate on money demand. The long-term interest rate supposedly captures household substitutions out of M2 and into long-term financial assets. The evidence here indicates that a standard M2 demand regression augmented to include the bond rate spread can account for most of the "missing M2" since 1990 if the estimation includes the missing

The author wishes to thank Robert Hetzel, Roy Webb, and Alex Wolman for many helpful comments. The views expressed are those of the author and not necessarily those of the Federal Reserve Bank of Richmond or the Federal Reserve System.

[1] See Greenspan (1993). The issue of the stability of money demand is central in assessing M2's usefulness for formulating policy. If M2 weakens, policymakers have to determine whether this weakness has resulted from a shift in money demand or whether it indicates that the Fed has been supplying an inadequate amount of money to the economy. If it's the latter, weak M2 growth may portend weakness in the economy.

To remind readers, the current definition of M2 includes currency, demand deposits, other checkable deposits, savings deposits, small-denomination time deposits, retail money market mutual funds and overnight repurchase agreements and Eurodollar deposits.

M2 period. Furthermore, changes in the missing M2 are highly correlated with changes in household holdings of bond and stock mutual funds from 1990 to 1994. This evidence lends credence to the view that the steepening of the yield curve in the early 1990s encouraged households to substitute out of M2 and into other financial assets and that part of this missing M2 ended up in bond and stock mutual funds.

However, a few caveats suggest caution in interpreting the twin role of the long-term interest rate and the growth of the mutual fund industry in influencing money demand. One is that the bond rate has no predictive content for M2 demand in the pre-missing M2 period. And during the past two years, 1995 and 1996, actual M2 growth has been in line with that predicted by the money demand regression estimated with and without the bond rate. Hence, the result that the bond rate can account for the missing M2 from 1990 to 1994 is interesting, but it does not necessarily indicate the presence of the systematic influence of the yield curve on M2 demand. The other caveat is that household holdings of bond and stock mutual funds continued to increase in 1995 and 1996, and that increase has not come at the expense of weak M2 growth. In fact, the strong correlation noted above between the missing M2 and household holdings of bond and stock mutual funds disappears when post-'94 observations are included. This result indicates that changes in household holdings of bond and stock mutual funds do not necessarily imply instability in M2 demand.

Taken together, one interpretation of this evidence is that special factors, such as the unusual steepening of the yield curve in the early '90s and the increased availability and liquidity of mutual funds since then, caused the public to redirect part of savings balances from bank deposits to bond and stock mutual funds. Those factors probably have not changed the character of M2 demand beyond causing a one-time permanent shift in the level of M2 balances demanded by the public.[2] The result that the leftward shift in M2 demand ended two years ago should now be of interest to monetary policymakers.

The plan of this article is as follows. Section 1 presents the standard M2 demand regression and reviews the econometric evidence indicating the existence of the missing M2 since 1990. Section 2 presents an explanation of the missing M2 and Section 3 examines the role of the bond rate in explaining the missing M2. Section 4 contains concluding observations.

[2] Other special factors that have usually been cited are resolution of thrifts by the Resolution Trust Corporation; the credit crunch; the downsizing of consumer balances accomplished by using M2 balances to pay off debt; rising deposit insurance premiums and the imposition of new, high-capital standards for depositories (resulting in a decreasing proportion of intermediation through the traditional banking sector); and so on. But none of these other special factors offers a satisfactory explanation of the missing M2 from 1990 to 1994 as does the steepening of the yield curve. See Duca (1993), Darin and Hetzel (1994), and Feinman (1994) for a further discussion of these special factors.

1. A STANDARD M2 DEMAND EQUATION AND ITS PREDICTIVE FAILURE IN THE EARLY 1990s

An M2 Demand Model

The money demand model that underlies the empirical work here is in error-correction form and is reproduced below (Mehra 1991, 1992):

$$m_t = a_0 + a_1 y_t + a_2 (R - RM2)_t + U_t \text{ and} \tag{1}$$

$$\Delta m_t = b_0 + \sum_{s=1}^{n1} b_{1s} \Delta m_{t-s} + \sum_{s=1}^{n2} b_{2s} \Delta y_{t-s} + \sum_{s=0}^{n3} b_{3s} \Delta (R - RM)_{t-s} + \lambda U_{t-1} + \epsilon_t, \tag{2}$$

where m is real M2 balances; y is real GDP; R is a short-term nominal interest rate; $RM2$ is the own rate on M2; U and ϵ are the random disturbance terms; and Δ is the first-difference operator. All variables are in their natural logs except interest rates. Equation (1) is the long-run equilibrium M2 demand function and is standard in the sense that the public's demand for real M2 balances depends upon a scale variable measured by real GDP and an opportunity cost variable measured as the difference between a short-term nominal rate of interest and the own rate of return on M2. The parameter a_1 measures the long-run income elasticity and a_2 is the long-run opportunity cost parameter. Equation (2) is the short-run money demand equation, which is in a dynamic error-correction form. The parameter b_{is} $(I = 2, 3)$ measures short-run responses of real M2 to changes in income and opportunity cost variables. The parameter λ is the error-correction coefficient. It is assumed that if variables in (1) are nonstationary in levels, they are cointegrated (Engle and Granger 1987). The presence of the error-correction mechanism indicates that if actual real money balances are high relative to what the public wishes to hold ($U_{t-1} > 0$), then the public will be reducing its holdings of money balances. Hence the parameter λ that appears on U_{t-1} in (2) is negative.

The long- and short-run money demand equations given above can be estimated jointly. This is shown in (3), which is obtained by solving for U_{t-1} in (1) and substituting in (2) (Mehra 1992):

$$\Delta m_t = d_0 + \sum_{s=1}^{k} b_{1s} \Delta m_{t-s} + \sum_{s=1}^{k} b_{2s} \Delta y_{t-s} + \sum_{s=0}^{n3} b_{3s} \Delta (R - RM2)_{t-s} + d_1 m_{t-1} + d_2 y_{t-1} + d_3 (R - RM2)_{t-1} + \epsilon_t, \tag{3}$$

where $d_0 = b_0 - \lambda a_0$; $d_1 = \lambda$; $d_2 = -\lambda a_1$; and $d_3 = -\lambda a_2$. As can be seen, the long-term income elasticity can be recovered from the long-run part of the money demand equation (3), i.e., a_1 is d_2 divided by d_1. If the long-term

income elasticity is unity ($a_1 = 1$ in [1]), then this assumption implies the following restriction on the long-run part of equation (3):

$$d_1 + d_2 = 0. \tag{4}$$

Equation (4) says that coefficients that appear on y_{t-1} and m_{t-1} sum to zero. The short-run part of (3) yields another estimate of the long-term income elasticity, i.e., as $\left(\sum_{s=0}^{n2} b_{2s}\right) / \left(1 - \sum_{s=1}^{n1} b_{1s}\right)$. If the same scale variable appears in the long- and short-run parts of the model, then a convergence condition can be imposed on equation (3) to ensure that one gets the same point-estimate of the long-run scale elasticity. The convergence condition implies another restriction (5) on the short-run part of equation (3):

$$\left(\sum_{s=0}^{n2} b_{2s}\right) / \left(1 - \sum_{s=1}^{n1} b_{1s}\right) = 1. \tag{5}$$

Equivalently, (5) can be expressed as

$$\sum_{s=0}^{n2} b_{2s} + \sum_{s=1}^{n1} b_{1s} = 1.$$

That is, coefficients that appear on Δm_{t-s} and Δy_{t-s} in (3) sum to unity. Equation (3) can be estimated by ordinary least squares or by instrumental variables if income and/or opportunity cost variables are contemporaneously correlated with the disturbance term.

An Estimated Standard M2 Demand Regression: 1960Q4 to 1989Q4

Panel A in Table 1 presents results of estimating the standard money demand regression (3) over the pre-missing M2 period, 1960Q4 to 1989Q4. Regressions are estimated using the new, chain-weighted price and income data.[3,4] I present

[3] The empirical work here uses the quarterly data over the period 1959Q3 to 1996Q4. Variables that appear in (3) are measured as follows. Real money balances (m) are the log of nominal M2 deflated by the GDP deflator; scale variables are the logs of real GDP and real consumer spending. All income and price data used are chain-weighted. R is the four-to-six-month commercial paper rate; $RM2$ is the weighted average of the explicit rates paid on the components of M2. The bond rate ($R10$) used later is the nominal yield on ten-year Treasury bonds. The data on household holdings of bond and equity mutual funds is from the Board of Governors and is constructed by adding net assets of mutual funds but netting out institutional and IRA/Keogh balances (Collins and Edwards 1994).

[4] Instrumental variables are used to estimate money demand regressions. Instruments used are just lagged values of the right-hand side explanatory variables. Ordinary least squares are not used mainly out of concern for the simultaneity bias. Both procedures yield similar estimates of the long-run parameters, even though estimates of short-run parameters differ. The convergence condition is usually rejected if ordinary least squares are used, but that is not the case with instrumental variables. That result favors instrumental variables. Nevertheless, the Hausman statistic (Hausman 1978) that tests the hypothesis that ordinary least squares estimates of all parameters are identical to those using the instrumental procedure is small, indicating that simultaneity may not be a serious problem.

Table 1 Instrumental Variable Estimates of M2 Demand Regressions: 1960Q4 to 1989Q4

Regression A M2 Demand without the Bond Rate

$$\Delta m_t = \underset{(4.3)}{-0.05} + \underset{(2.9)}{0.23}\ \Delta m_{t-1} + \underset{(1.0)}{0.08}\ \Delta m_{t-2} + \underset{(4.3)}{0.45}\ \Delta c_t + \underset{(3.6)}{0.24}\ \Delta c_{t-1}$$

$$\underset{(1.6)}{-0.002}\ \Delta(R - RM2)_t - \underset{(3.7)}{0.003}\ \Delta(R - RM2)_{t-1} - \underset{(4.6)}{0.11}\ m_{t-1} + \underset{(4.6)}{0.11}\ \tilde{y}_{t-1}$$

$$\underset{(3.8)}{-0.002}\ (R - RM2)_{t-1} - \underset{(4.2)}{0.72}\ T_t + \underset{(5.5)}{0.03}\ D83Q1$$

$CRSQ = 0.78 \quad SER = 0.0047 \quad Q(2) = 1.5 \quad Q(4) = 5.1 \quad Q(29) = 22.6$

$N_c = N_y = 1.0 \quad L_{(R-RM2)} = -0.02$

$F1(2,105) = 0.99$

Regression B M2 Demand with the Bond Rate

$$\Delta m_t = \underset{(4.2)}{-0.05} + \underset{(3.5)}{0.26}\ \Delta m_{t-1} + \underset{(1.1)}{0.08}\ \Delta m_{t-2} + \underset{(4.2)}{0.40}\ \Delta c_t + \underset{(4.0)}{0.26}\ \Delta c_{t-1}$$

$$\underset{(1.5)}{-0.002}\ \Delta(R - RM2)_t - \underset{(4.0)}{0.004}\ \Delta(R - RM2)_{t-1} - \underset{(4.4)}{0.11}\ m_{t-1} + \underset{(4.4)}{0.11}\ \tilde{y}_{t-1}$$

$$\underset{(3.1)}{-0.002}\ (R - RM2)_{t-1} - \underset{(3.3)}{0.63}\ T_t + \underset{(5.2)}{0.03}\ D83Q1 - \underset{(0.7)}{0.005}\ (R10 - RM2)_{t-1}$$

$$\underset{(1.5)}{-0.002}\Delta(R10 - RM2)_{t-1}$$

$CRSQ = 0.79 \quad SER = 0.0045 \quad Q(2) = 1.5 \quad Q(4) = 5.2 \quad Q(29) = 26.9$

$N_c = N_y = 1.0 \quad L_{(R-RM2)} = -0.02 \quad L_{(R10-RM2)} = -0.004$

$F1(2,105) = 0.99 \quad F2(2,105) = 2.24$

Notes: m is real M2 balances; c is real consumer spending; $\tilde{y}$ is $(y_t + y_{t-1})/2$ where y is real GDP; R is the four-to-six-month commercial paper rate; $RM2$ is the own rate on M2; $R10$ is the nominal yield on ten-year U.S. Treasury bonds; $D83Q1$ is a dummy that equals 1 in 83Q1 and 0 otherwise; Δ is the first difference operator. All variables are in their natural logs, except interest rate variables. $CRSQ$ is the corrected R-squared; SER is the standard error of regression; $Q(k)$ the Ljung-Box Q-statistic based on k number of auto correlations of the residuals. N_y is the long-term income elasticity; N_c is the long-term consumption elasticity; $N_{(R-RM2)}$ is the long-term opportunity cost parameter. $F1$ tests the restriction $N_y = N_c = 1$; $F2$ tests the restriction that the bond rate spread variables are not significant in the regression (the 5 percent critical value is 3.1). Instruments used for estimation are just lagged values of the right-hand side explanatory variables. The reported coefficient on trend is to be divided by 1,000.

the version estimated using real consumer spending as the short-run scale variable and real GDP as the long-run scale variable. The evidence reported in Mankiw and Summers (1986), Small and Porter (1989), and Mehra (1992) indicates that in the short run changes in real money balances are correlated more with changes in consumer spending than with real GDP.[5] The regression, however, is estimated under the assumption that the long-run scale elasticity is unity, computed using either the long-run part or the short-run part of (3). That is, restrictions (4) and (5) are imposed on equation (3). In addition, the regression includes a deterministic time trend and a dummy for the introduction of superNows and money market deposit accounts.[6]

As can be seen, the coefficients that appear on the scale and opportunity cost variables have theoretically correct signs and are statistically significant.[7] F1 tests the restrictions that long-run income and consumer spending elasticities are unity. This F-statistic is small, indicating that those restrictions are consistent with data (see Table 1). The long-run opportunity cost parameter is -0.02, indicating that a 1 percentage point increase in M2's opportunity cost ($R - RM2$) from its current level would reduce equilibrium M2 demand by about 2 percent. It is also worth noting that the long-run part of the money demand equation is well estimated. In particular, the estimated error-correction coefficient is correctly signed and significant, indicating the presence of a cointegrating M2 relation in the pre-1990 period.

Evidence on the Missing M2 during the 1990s

Panel A in Table 2 presents the dynamic, out-of-sample predictions of M2 growth from 1990Q1 to 1996Q4. Those predictions are generated using the standard M2 demand regression given in Table 1. Actual M2 growth and prediction errors (with summary statistics) are also reported. As shown in the

[5] I prefer to work with this specification because the restrictions that the long-run scale elasticity computed using either the long-run part or the short-run part is unity are consistent with the data in this specification. Those restrictions are usually found to be inconsistent with the data when instead real GDP is used in the short-run part. Nevertheless, the results here are not sensitive to the use of different scale variables in short- and long-run parts of the money demand equation. In particular, with real GDP in the short-run part we still have the episode of the missing M2 from 1990 to 1994 and the result that M2 growth was on track in the years 1995 and 1996.

[6] In the empirical money demand literature, time-trend variables generally proxy for the effect of ongoing financial innovation on the demand for money. Estimates reported in many previous studies indicate that the statistical significance of trend variables in money demand regressions is not robust across different specifications and sample periods. For example, a time trend when included in the Federal Reserve Board M2 demand model is significant (Small and Porter 1989; Duca 1995; Koenig 1996), whereas that is not the case in specifications reported in Hetzel and Mehra (1989) and Mehra (1991, 1992). Different sample periods used in these studies may account for these different results.

[7] The Ljung-Box Q-statistics presented in Table 1 indicate that serial correlation is not a problem.

Table 2 Evidence on Missing M2 during the 1990s

Year	Actual M2 Growth	Panel A Regression A: Predicted M2 Growth	Panel A Regression A: Error			Panel B Regression B: Predicted M2 Growth	Panel B Regression B: Error		
			Growth	Cumulative			Growth	Cumulative	
				Level (billions)	Percentage			Level (billions)	Percentage
1990Q4	4.0	6.4	−2.3	−71	2.2	6.5	−2.4	−80	2.4
1991Q4	3.0	3.6	−0.5	−91	2.7	3.3	−0.3	−92	2.7
1992Q4	1.8	6.4	−4.5	−257	7.5	5.9	−4.0	−239	6.9
1993Q4	1.4	4.8	−3.4	−392	11.2	5.0	−3.6	−381	10.9
1994Q4	0.6	3.0	−2.4	−489	13.9	2.6	−2.0	−464	13.2
1995Q4	3.8	3.5	0.3	−495	13.6	4.2	−0.4	−500	13.7
1996Q4	4.5	3.9	0.5	−495	13.0	4.0	−0.4	−505	13.3
Mean Error (1990–1996)			−1.78				−1.78		
RMSE			2.52				2.40		

Notes: The predicted values are generated using the regressions reported in Table 1. Regressions are estimated from 1960Q4 to 1989Q4 and dynamically simulated from 1990Q1 to 1996Q4. *RMSE* is the root mean squared error.

table, this money demand regression overpredicts M2 growth from 1990 to 1994. Those prediction errors cumulate to an overprediction in the level of M2 of about $490 billion, or 14 percent, by the fourth quarter of 1994.[8]

However, since 1995 M2 growth has been in line with that predicted by the money demand regression. The cumulative over prediction in the level of M2 has stabilized and there is no tendency for the level percent error to increase since then (see panel A in Table 2). This evidence indicates that the leftward shift in the public's M2 demand seen early in the 1990s may have ended.

2. AN EXPLANATION OF THE MISSING M2

Portfolio-Substitution Hypothesis

It is widely held that weak M2 growth observed in the early '90s is due to household substitutions out of bank deposits (in M2) and into long-term financial assets including bond and stock mutual funds.[9] Two developments may have contributed to such portfolio substitution. One is the increased availability and liquidity of bond and stock mutual funds brought about by reductions in transaction costs, improvements in computer technology, and the introduction of check writing on mutual funds. The other is the steepening of the yield curve brought about mainly by a reduction in short-term market interest rates in general and bank deposit rates in particular.[10] It is suggested that the combination of these factors reduced the public's demand for savings in the form of bank deposits, leading them to redirect savings balances into long-term financial assets including bond and stock mutual funds.[11]

[8] This predictive failure is confirmed by formal tests of stability. The conventional Chow test with the shift date (1978Q4) located near the midpoint of the sample period indicates that the M2 demand regression is unstable from 1960Q4 to 1996Q4. The Dufour test (Dufour 1980), which is a variant of the Chow test, examines stability over the particular interval, 1990Q1 to 1994Q4. This test uses an F-statistic to test the joint-significance of dummy variables introduced for each observation over 1990Q1 to 1994Q4. The results here indicate that the individual coefficients that appear on these shift dummies are generally large and statistically significant. The F-statistic is large and significant at the 10 percent level. (The F-statistic, however, is not significant at the 5 percent level.) Together these results indicate that the M2 demand regression is not stable over this interval.

[9] Darin and Hetzel (1994), Wenninger and Partlan (1992), Feinman and Porter (1992), Collins and Edwards (1994), Orphanides, Reid, and Small (1994), Duca (1995), and Koenig (1996). Wenninger and Partlan (1992) argued that weakness in M2 growth was due to weakness in its small time deposits component.

[10] Many analysts have argued that the decline in the size of taxpayers' subsidy to the depository sector also may have contributed to a reduction in offering rates on bank deposits. It is argued that rising premiums for deposit insurance, higher capital requirements, and more stringent standards for depository borrowing and lending in both wholesale and retail markets may have pressured many banks and thrifts to widen intermediation margins, resulting in lower offering rates on many bank deposits (Feinman and Porter 1992).

[11] It may, however, be noted that bond and stock funds also grew rapidly in the mid '80s, shortly after IRA, 401k, and Keogh regulations were liberalized. Such growth, however, did not

Tests in Previous Studies

The portfolio-substitution hypothesis outlined above has been tested in two different ways. The first one attempts to internalize such substitutions by adding bond and/or stock mutual funds to M2. Duca (1995) adds bond funds to M2 and finds the expanded M2 more explainable from 1990Q3 to 1992Q4. Darin and Hetzel (1994) shift-adjust M2, and Orphanides, Reid, and Small (1994) simply add bond and stock funds to M2. While the resulting monetary aggregates do explain part of the missing M2 or improve the predictive content of M2 in the missing M2 period, they worsen performance in other periods.[12]

The other approach attempts to capture the increased substitution of mutual funds for bank deposits by redefining the opportunity cost of M2 to include the long-term bond rate. This approach assumes that the bond rate is a proxy for the return available on long-term financial assets including bond and stock mutual funds. Hence M2 demand is assumed to be sensitive to both short- and long-term interest rates (Feinman and Porter 1992; Mehra 1992; Koenig 1996). This approach has been relatively more successful in explaining the missing M2 than the other one discussed above.

The main issue here however is whether the character of M2 demand has changed since 1990. In Koenig (1996) long-term interest rates are found to influence M2 demand even before the period of missing money, suggesting that the character of M2 demand did not change and that standard M2 demand regressions estimated without the long-term interest rate are misspecified. In contrast the empirical work in Feinman and Porter (1992) and Mehra (1992) are consistent with the observation that long-term interest rates did not add much towards explaining M2 demand in pre-1990 sample periods. In the next section I examine further the quantitative importance of the long-term interest rate in explaining M2 demand.

3. THE ROLE OF THE BOND RATE IN M2 DEMAND

Pre-1990 M2 Demand Regression with the Bond Rate

Panel B in Table 1 presents the standard M2 demand regression augmented to include the bond rate spread variable measured as the difference between the nominal yield on ten-year Treasury bonds and the own rate of return on M2.

destablize M2 demand. The flow-of-funds data discussed in Duca (1994) indicates that the assets that households shifted into bond and equity funds came from direct holdings of bonds and equities rather than from M2 deposits. By contrast more of the inflows into bond and stock funds in the early ’90s reflected shifts out of M2 rather than out of direct bond and equity holdings.

[12] For example, Orphanides, Reid, and Small (1994) report that money demand equations that add bond and stock funds to M2 fail Chow tests of stability. Koenig (1996) shows that the bond-fund adjusted M2 demand equation, while it improved the forecast performance from 1990 to 1994, worsened performance in the early sample period. I show later (see footnote 16) that adding bond and stock funds to M2 worsened performance over the last couple of years.

I include both the level and first differences of this spread. The regression is estimated over the pre-missing M2 demand period, 1960Q4 to 1989Q4. It is evident that the coefficient that appears on the level of the bond rate spread variable is small and statistically not different from zero. F2 is the F-statistic that tests the hypothesis that coefficients that appear on both the level and first differences of the bond rate spread variable are zero. This statistic is small, indicating that the bond rate spread did not influence M2 demand in the pre-1990 period (see regression B in Table 1).

Including the bond rate spread in the M2 demand regression estimated using only pre-1990 sample observations does not solve the missing M2 puzzle either. The evidence on this point is indicated by the dynamic out-of-sample simulations of M2 demand given in panel B of Table 2. The augmented M2 demand regression continues to overpredict M2 growth from 1990 to 1994. Those prediction errors cumulate to an overprediction in the level of M2 of about $464 billion, or 13.2 percent by end of 1994. Including the bond rate spread does yield a somewhat lower root mean squared error, but this improvement is very small (compare prediction errors in panels A and B of Table 2).[13]

Full-Sample M2 Demand Regression with the Bond Rate

Table 3 presents M2 demand regressions estimated including post-'90 sample observations. In regression D the bond rate spread enters interacting with a slope dummy that is unity since 1989 and zero otherwise. In that specification the restriction that the bond rate spread did not influence M2 demand in the pre-1990 period is imposed on the regression. I also present the regression C

[13] In the money demand regression above, long- and short-rate spreads are included in an unrestricted fashion. It is possible to get the result that the bond rate influenced M2 demand even before the missing M2 demand period if the opportunity cost of holding M2 is alternatively measured as a weighted average of long- and short-rates:

$$OC_t = (w * R10_t + (1 - w) * RCP_t) - RM2_t,$$

where OC_t is the opportunity cost; w is the weighting coefficient; and other variables are defined as before. If $w = 0$, then the bond rate is not relevant in influencing M2 demand.

The money demand regression (3) here also is estimated using this alternative measure. Estimation results using pre-1990 sample observations indicate that the standard error of M2 demand regression is minimized when $w = 0.4$. In that regression the opportunity cost variable is correctly signed and significant, indicating that the long-term interest rate influenced M2 demand even before the missing M2 demand period. This finding is similar to the one reported in Koenig (1996). However, this empirical specification does not solve the missing M2 problem. M2 growth predicted by this regression remains large relative to actual M2 growth from 1990 to 1994. Those prediction errors still cumulate to generate an overprediction in the level of M2 of about $441 billion, or 12.6 percent by end of 1994. The magnitude of this prediction error is somewhat smaller than the one generated by assuming $w = 0$. But the improvement is small. This empirical specification does not solve the missing M2 problem because the increased explanatory power of the bond rate in the M2 demand regression comes at the cost of the short rate.

Table 3 Instrumental Variables Estimates of M2 Demand Regressions: 1960Q4 to 1996Q4

Regression C M2 Demand with the Bond Rate, but No Slope Dummies

$$\begin{aligned}\Delta m_t = &-0.01 + 0.33\ \Delta m_{t-1} + 0.13\ \Delta m_{t-2} + 0.36\Delta c_t + 0.17\ \Delta c_{t-1}\\ &(2.1)\quad (4.8)\qquad (1.9)\qquad (3.9)\qquad (2.7)\\ &-0.001\Delta(R-RM2)_t - 0.005\ \Delta(R-RM2)_{t-1} - 0.03\ m_{t-1} + 0.03\ \tilde{y}_{t-1}\\ &(0.9)\qquad (5.9)\qquad (2.4)\qquad (2.4)\\ &-0.000\ (R-RM2)_{t-1} - 0.000\ (R10-RM2)_{t-1} - 0.004\ \Delta(R10-RM2)_{t-1}\\ &(0.3)\qquad (0.1)\qquad (3.2)\\ &-0.51\ T_t + 0.02\ D83Q1\\ &(3.0)\qquad (5.1)\end{aligned}$$

$CRSQ = 0.78 \quad SER = 0.0047 \quad Q(2) = 2.7 \quad Q(4) = 6.6 \quad Q(29) = 35.1$

$N_y = N_c = 1 \quad N_{(R-RM2)} = -0.005 \quad N_{(R10-RM2)} = -0.005$

$F2(2,133) = 5.5^*$

Regression D M2 Demand with the Bond Rate Interacting with Slope Dummy

$$\begin{aligned}\Delta m_t = &-0.03 + 0.25\ \Delta m_{t-1} + 0.08\ \Delta m_{t-2} + 0.49\ \Delta c_t + 0.18\ \Delta c_{t-1}\\ &(3.8)\quad (3.5)\qquad (1.1)\qquad (5.2)\qquad (2.8)\\ &-0.002\ \Delta(R-RM2)_t - 0.003\ \Delta(R-RM2)_{t-1} - 0.07\ m_{t-1} + 0.07\tilde{y}_{t-1}\\ &(1.5)\qquad (4.2)\qquad (3.9)\qquad (3.9)\\ &-0.001(R-RM2)_{t-1} - 0.002\ (D*R10-RM2)_{t-1} - 0.003\ \Delta(R10-RM2)_{t-1}\\ &(2.5)\qquad (3.2)\qquad (2.1)\\ &-0.56\ T_t + 0.02\ D83Q1\\ &(3.6)\qquad (5.3)\end{aligned}$$

$CRSQ = 0.77 \quad SER = 0.0046 \quad Q(2) = 1.7 \quad Q(4) = 6.0 \quad Q(36) = 35.3$

$N_y = N_c = 1 \quad N_{(R-RM2)} = -0.02 \quad N_{(R10-RM2)} = -0.03$

*Significant at the 5 percent level.

Notes: D is a dummy that is 1 from 1989Q1 to 1996Q4 and 0 otherwise. $N_{(R10-RM2)}$ is the long-run bond rate opportunity cost parameter. See also notes in Table 1.

in which no such slope dummy is included. Both differences and the level of the bond rate spread are included in these regressions. As can be seen, the coefficient that appears on the level of the bond rate spread is significant

only in the regression where the spread is included interacting with the slope dummy.[14] Moreover, in that regression other coefficients, including the one that appears on the error-correction variable, have expected signs and are statistically significant. In contrast none of the coefficients that appear on levels of the interest rate spreads are significant in the regression without the slope dummy (compare coefficients in regressions C and D of Table 3).[15] Together this evidence indicates that a significant role for the impact of the long-term interest rate on M2 demand emerges only in the post-1990 period.[16]

Panel D in Table 4 presents the dynamic, within-sample simulations of M2 growth from 1990 to 1996 generated using the regression with the slope dummy. As shown in the table, this regression can account for most of the missing M2 since 1990. The prediction errors now cumulate to an overprediction in the level of M2 of about $41 billion, or 1.2 percent by end of 1994. Since then, the level percent error has displayed no tendency to increase over time. This

[14] The intuition behind this result is that the least squares regression coefficient measures the average response of M2 demand to the spread variables over the full sample. If for most of the sample period—as is the case here—this response is small or zero, then the estimated regression coefficient that simplify averages such responses over the full sample will be small or zero. But when the slope dummy is included, the estimated regression coefficient receives full weight over part of the sample over which the response is believed to be strong.

I have not reported the slope dummy on the first difference of the bond rate spread because it is not significant in the regression.

[15] In the regression C without the slope dummy, the error-correction coefficient is small in magnitude and only marginally significant. In fact, if restrictions that long-run scale elasticities are unity are not imposed on the regression, then none of the coefficients that appear on levels of variables are significant. Hence in these regressions the hypothesis that there exists a cointegrating M2 demand relation is easily rejected. This finding is similar in spirit to the one in Miyao (1996), where it is shown that once post-'90 sample observations are included in the estimation period, evidence supports no M2 cointegration.

[16] Alternatively, the hypothesis that most of the missing M2 went into bond and stock mutual funds can be tested by broadening the definition of M2 to include such mutual funds. If the hypothesis is correct, then the broadly defined monetary aggregate should be more explainable from 1990 to 1994. This procedure yields similar results. To explain it further, consider the behavior of the monetary aggregate that simply adds bond and stock mutual funds to M2, denoted hereafter as M2+ (Orphanides, Reid, and Small 1994). This aggregate has grown at the following rates (in percent) in recent years: 4.1 in 1990, 6.2 in 1991, 4.6 in 1992, 5.5 in 1993, 0.9 in 1994, 6.3 in 1995, and 7.9 in 1996. For those years M2 growth predicted by the standard M2 demand regression is 6.4, 3.5, 6.4, 4.8, 3.0, 3.5, and 3.9, respectively. The corresponding prediction errors are -2.3, 2.0, -1.7, 0.6, -2.0, 2.9, and 3.9. As can be easily verified, for the period 1990 to 1994 the mean prediction error is -0.57 percentage point and the root mean squared error is 2.0 percentage points. These prediction errors are smaller than those generated using the narrowly defined M2; for the latter the mean error is -1.78 and the root mean squared error is 2.52. Thus M2+ is more explainable over the period 1990 to 1994 than is M2. However, adding bond and stock funds to M2 does not yield a more stable money demand equation. As can be seen, strong growth in M2+ over the period 1995 to 1996 is not easily predicted when conventional money demand parameters are used to characterize M2+ demand. The analysis above, however, is subject to the caveat that the opportunity cost variable in M2+ demand is different from the one that shows up in M2 demand. In particular, the own rate of return on M2+ must include the returns on bond and stock mutual funds.

Table 4 Role of the Bond Rate in Explaining the Missing M2 during the 1990s

		Panel D Regression D				Panel E	
Year	Actual M2 Growth	Predicted M2 Growth	Error			Missing M2[a] Explained by the Bond Rate	Cumulative Change since 1989 in Household Holdings of Bond and Stock Mutual Funds[b]
			Growth	Cumulative			
				Level	Percentage		
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1990	4.0	3.6	0.4	22	0.7	68	2
1991	3.0	0.5	2.5	107	3.2	163	109
1992	1.8	3.4	−1.5	56	1.6	265	212
1993	1.4	3.3	−1.9	11	−0.3	327	370
1994	0.6	1.4	−0.8	−41	1.2	404	386
1995	3.8	3.5	0.3	−31	−0.8	421	505
1996	4.5	4.3	−0.2	−24	−0.7	422	680
Mean (1990–1996)			−0.13				
RMSE			1.41				

[a]in billions

[b]Household holdings are net of institutional and IRA/Keogh assets (Collins and Edwards 1994).

Notes: The predicted values in panel D are the within-sample dynamic simulations generated using M2 demand regression D reported in Table 3.

Column (7) above reports the missing M2 that is predicted by the inclusion of the bond rate in M2 demand regression D. These values are generated by comparing predictions of M2 demand with and without including the bond rate.

evidence indicates that the steepening of the yield curve contributed to weak M2 growth in the early '90s.

The Missing M2 and Bond and Stock Mutual Funds

Figure 1 charts the missing M2 as explained by the bond rate spread since 1990.[17] It also charts the cumulative change (since 1989) in household holdings of bond and stock mutual funds.[18] As can be seen, these two series comove from 1990 to 1994. But this comovement ends in the years 1995 and 1996. Furthermore, in the beginning years (1990, 1991, 1992) of this missing period, the magnitude of the missing M2 somewhat exceeds the cumulative increase in household holdings of bond and stock mutual funds. This data supports the view that weak M2 growth in the early '90s is due to household's substitution out of M2 and into bond and stock mutual funds. But not all of the missing M2 first went into bond and stock funds. A part might have gone into direct holdings of bonds, stocks, and other long-term savings vehicles (Duca 1993; Darin and Hetzel 1994).

If part of the missing M2 ended up in bond and stock mutual funds, then changes in missing M2 balances should be correlated with changes in household holdings of bond and stock funds. This implication is tested by running the following regression:

$$\Delta BS_t = a_0 + a_1 \Delta BS_{t-1} + a_2 \Delta MM2_t + \epsilon_t,$$

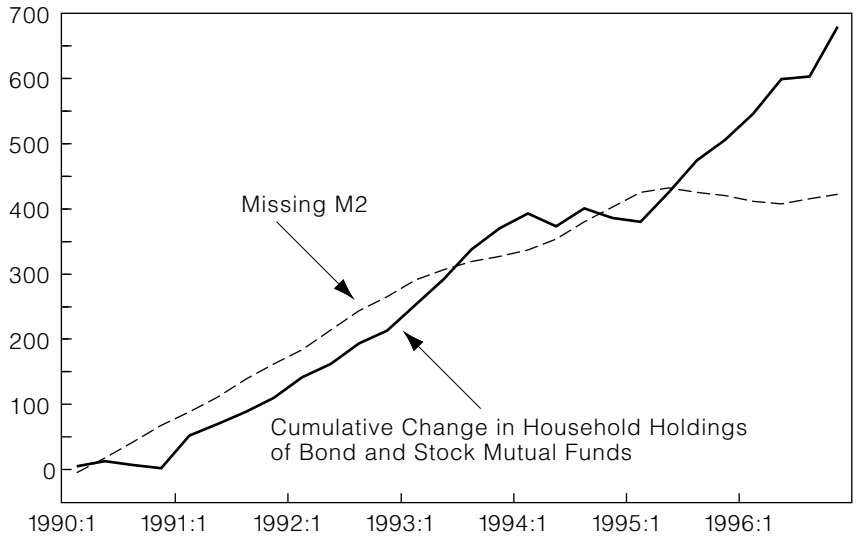
where *BS* is household holdings of bond and stock funds; *MM*2 is the missing M2; and ϵ_t is the random disturbance term. The series on *BS* and *MM*2 are reported in Table 4 and charted in Figure 1. Estimation results indicate that from 1991Q1 to 1994Q4 $a_2 \neq 0$, but from 1991Q1 to 1996Q4 $a_2 = 0$.[19] These results are consistent with the hypothesis that part of the missing M2 during the 1990s ended up in bond and stock mutual funds.

[17] The series on the missing M2 is generated in the following way. The M2 demand regression D, which includes the bond rate interacting with a slope dummy, is estimated from 1960Q4 to 1996Q4. This regression is dynamically simulated from 1990Q1 to 1996Q4, first using actual values of the bond rate spread over the prediction interval and then repeating the simulation with actual values of the bond rate set to zero. The difference in predicted values so generated gives M2 demand explained by the bond rate.

[18] This series is constructed by Collins and Edwards (1994) and is the plus part of the monetary aggregate (M2+) discussed in the previous footnote. As noted before, the plus part is the market value of household holdings of bond and stock mutual funds. The current definition of the conventional M2 aggregate includes currency, demand deposits, other checkable deposits, savings deposits, small time deposits, retail money market mutual funds and overnight RPs, and Eurodollar deposits. Since this definition does not include institutional and IRA/Keogh balances, household holdings of bond and stock funds are also net of such assets. However, unlike M2, those household holdings can increase if bonds and stocks appreciate and thus do not necessarily represent funds out of new savings.

[19] The regressions use quarterly observations on year-over-year changes in *BS* and *MM*2 and are run from 1991Q1 to 1996Q4.

Figure 1 The Missing M2 and the Cumulative Change in Household Holdings of Bond and Stock Mutual Funds since 1990



Notes: The missing M2 is the reduction in M2 demand that is due to the bond rate spread. Household holdings are net of institutional and IRA/Keogh assets.

4. CONCLUDING OBSERVATIONS

It is now known that the public's demand for M2 experienced a leftward shift in the early '90s. It is widely believed that this shift reflected the public's desire to redirect savings balances from bank deposits to long-term financial assets, including bond and stock mutual funds. In this article, I test this popular hypothesis. In particular, I present evidence that a standard M2 demand regression augmented to capture the impact of the long-term interest rate on money demand can account for most of the missing M2 since 1990 and that changes in this missing M2 are highly correlated with changes in household holdings of bond and stock mutual funds in the early 1990s.

The evidence here, however, also indicates that the long-term interest rate has no predictive content for M2 demand in the pre-missing M2 period. That result suggests caution in assigning a causal role to the independent influence of the long-term rate on M2 demand found in the missing M2 period. Furthermore, household holdings of bond and stock mutual funds continued to increase in the years 1995 and 1996, but that increase has not accompanied any weakness

in M2. Hence increases in household holdings of bond and stock mutual funds may not necessarily signal instability in M2 demand.

One interpretation of the recent behavior of M2 demand is that some special factors caused a leftward shift in the public's M2 demand. The evidence here is consistent with the view that those special factors included the combination of the unusual steepening of the yield curve and the increased availability, liquidity, and public awareness of bond and stock mutual funds. The evidence so far is that those special factors have not changed fundamentally the character of M2 demand beyond causing a one-time permanent shift in the level of M2 balances demanded by the public. Hence the result that the leftward shift in M2 demand ended two years ago should now be of interest to monetary policymakers.

REFERENCES

Collins, Sean, and Cheryl L. Edwards. "An Alternative Monetary Aggregate: M2 Plus Household Holdings of Bond and Equity Mutual Funds," Federal Reserve Bank of St. Louis *Review,* vol. 76 (November/December 1994), pp. 7–29.

Darin, Robert, and Robert L. Hetzel. "A Shift-Adjusted M2 Indicator of Monetary Policy," Federal Reserve Bank of Richmond *Economic Quarterly,* vol. 80 (Summer 1994), pp. 25–47.

Duca, John V. "Should Bond Funds be Added to M2?" *Journal of Banking and Finance,* vol. 19 (April 1995), pp. 131–52.

_____. "Commentary," Federal Reserve Bank of St. Louis *Review,* vol. 76 (November/December 1994), pp. 67–70.

_____. "RTC Activity and the 'Missing M2,'" *Economic Letters,* vol. 41 (1993), pp. 67–71.

Dufour, Jean Marie. "Dummy Variables and Predictive Tests for Structural Change," *Economic Letters,* vol. 6 (1980), pp. 241–47.

Engle, Robert F., and C. W. J. Granger. "Co-Integration and Error Correction: Representation, Estimation, and Testing," *Econometrica,* vol. 55 (March 1987), pp. 251–76.

Feinman, Joshua. "Commentary," Federal Reserve Bank of St. Louis *Review,* vol. 76 (November/December 1994), pp. 71–73.

_____, and Richard D. Porter. "The Continuing Weakness in M2," Finance and Economics Discussion Working Paper 209. Board of Governors of the Federal Reserve System, 1992, pp. 1–41.

Greenspan, Alan. "Statement to the Congress," *Federal Reserve Bulletin*, vol. 79 (September 1993), pp. 849–55.

Hausman, J. A. "Specification Tests in Econometrics," *Econometrica*, vol. 46 (November 1978), pp. 1251–72.

Hetzel, Robert L., and Yash P. Mehra. "The Behavior of Money Demand in the 1980s," *Journal of Money, Credit, and Banking*, vol. 21 (November 1989), pp. 455–63.

Koenig, Evan F. "Long-Term Interest Rates and the Recent Weakness in M2," *Journal of Economics and Business*, vol. 48 (May 1996), pp. 81–101.

Mankiw, N. Gregory, and Lawrence H. Summers. "Money Demand and the Effects of Fiscal Policies," *Journal of Money, Credit, and Banking*, vol. 18 (November 1986), pp. 415–29.

Mehra, Yash P. "Has M2 Demand Become Unstable?" Federal Reserve Bank of Richmond *Economic Review*, vol. 78 (September/October 1992), pp. 27–35.

_____. "An Error-Correction Model of U.S. M2 Demand," Federal Reserve Bank of Richmond *Economic Review*, vol. 77 (May/June 1991), pp. 3–12.

Miyao, Ryuzo. "Does a Cointegrating M2 Demand Relation Really Exist in the United States?" *Journal of Money, Credit, and Banking*, vol. 28 (August 1996), pp. 365–80.

Orphanides, Athanasios, Brian Reid, and David H. Small. "The Empirical Properties of a Monetary Aggregate that Adds Bond and Stock Funds to M2," Federal Reserve Bank of St. Louis *Review*, vol. 76 (November/December 1994), pp. 31–51.

Small, David H., and Richard D. Porter. "Understanding the Behavior of M2 and V2," *Federal Reserve Bulletin*, vol. 75 (April 1989), pp. 244–54.

Wenninger, John, and John Partlan. "Small Time Deposits and the Recent Weakness in M2," Federal Reserve Bank of New York *Quarterly Review*, vol. 17 (Spring 1992), pp. 21–35.

On the Identification of Structural Vector Autoregressions

Pierre-Daniel G. Sarte

Following seminal work by Sims (1980a, 1980b), the economics profession has become increasingly concerned with studying sources of economic fluctuations. Sims's use of vector autoregressions (VARs) made it possible to address both the relative importance and the dynamic effect of various shocks on macroeconomic variables. This type of empirical analysis has had at least two important consequences. First, by deepening policymakers' understanding of how economic variables respond to demand versus supply shocks, it has enabled them to better respond to a constantly changing environment. Second, VARs have become especially useful in guiding macroeconomists towards building structural models that are more consistent with the data.

According to Sims (1980b), VARs simply represented an atheoretical technique for describing how a set of historical data was generated by random innovations in the variables of interest. This reduced-form interpretation of VARs, however, was strongly criticized by Cooley and Leroy (1985), as well as by Bernanke (1986). At the heart of the critique lies the observation that VAR results cannot be interpreted independently of a more structural macroeconomic model. Recovering the structural parameters from an estimation procedure requires that some restrictions be imposed. These are known as identifying restrictions. Implicitly, the choice of variable ordering in a reduced-form VAR constitutes such an identifying restriction.

As a result of the Cooley-Leroy/Bernanke critique, economists began to focus more precisely upon the issue of identifying restrictions. The extent to which specific innovations were allowed to affect some subset of variables,

I would like to thank Tom Cooley, Michael Dotsey, Bruce Hansen, Tom Humphrey, Yash Mehra, and Alex Wolman for more than helpful comments. I would also like to thank Sergio Rebelo, Vassilios Patikis, and Mark Watson for their suggestions. The opinions expressed herein are the author's and do not represent those of the Federal Reserve Bank of Richmond or the Federal Reserve System.

either in the short run or in the long run, began to be derived explicitly from structural macroeconomic models. Consequently, what were previously considered random surprises could be interpreted in terms of specific shocks, such as technology or fiscal policy shocks. This more refined use of VARs, known as structural vector autoregressions (SVARs), has become a popular tool for evaluating economic models, particularly in the macroeconomics literature.

The fact that nontrivial restrictions must be imposed for SVARs to be identified suggests, at least in principle, that estimation results may be contingent on the choice of restrictions. To take a concrete and recent example, in estimating a system containing employment and productivity variables, Gali (1996) achieves identification by assuming that aggregate demand shocks do not affect productivity in the long run. Using postwar U.S. data, he is then able to show that, surprisingly, employment responds negatively to a positive technology shock. One may wonder, however, whether his results would change significantly under alternative restrictions. This article consequently investigates how the use of different identifying restrictions affects empirical evidence about business fluctuations. Two important conclusions emerge from the analysis.

First, by thinking of SVARs within the framework of instrumental variables estimation, it will become clear that the method is inappropriate for certain identifying restrictions. This finding occurs because SVARs use the estimated residual from a previous equation in the system as an instrument in the current equation. Since estimation of this residual depends on some prior identifying restriction, the identification scheme necessarily determines the strength of the instrument. By drawing from the literature on estimation with weak instruments, this article points out that in some cases, SVARs will not yield meaningful parameter estimates.

The second finding of interest suggests that even in cases where SVAR parameters can be properly estimated, different identification choices can lead to contradictory results. For example, in Gali (1996) the restriction that aggregate demand shocks not affect productivity in the long run also implies that employment responds negatively to a positive technology shock. But the opposite result emerges when aggregate demand shocks are allowed to have a small negative effect on productivity in the long run. This latter restriction is appropriate if demand shocks are interpreted as fiscal policy shocks in a real business cycle model. More importantly, this observation suggests that sensitivity analysis should form an integral part of deciding what constitutes a stylized fact within the confines of SVAR estimation.

This article is organized as follows. We first provide a brief description of reduced-form VARs as well as the basic idea underlying the Cooley-Leroy/Bernanke critique. In doing so, the important assumptions underlying the use of VARs are laid out explicitly for the nonspecialist reader. We then introduce the mechanics of SVARs—that is, the details of how SVARs are usually estimated—and link the issue of identification to the estimation

procedure.[1] The next section draws from the literature on instrumental variables in order to show the conditions in which the SVAR methodology fails to yield meaningful parameter estimates. We then describe the type of interpretational ambiguities that may arise when the same SVAR is estimated using alternative identifying restrictions. Finally, we offer a brief summary and some conclusions.

1. REDUCED-FORM VARs AND THE COOLEY-LEROY/BERNANKE CRITIQUE

In this section, we briefly describe the VAR approach first advocated by Sims (1980a, 1980b). In doing so, we will show that the issue of identification already emerges in interpreting estimated dynamic responses for a given set of variables. To make matters more concrete, the analysis in both this and the next section is framed within the context of a generic bivariate system. However, the basic issues under consideration are invariant with respect to the size of the system. Thus, consider the joint time series behavior of the vector $(\Delta y_t, \Delta x_t)$, which we summarize as

$$B(L)Y_t = e_t, \text{ with } B(0) = B_0 = I, \tag{1}$$

where $Y_t = (\Delta y_t, \Delta x_t)'$, and $B(L)$ denotes a matrix polynomial in the lag operator L. $B(L)$ is thus defined as $B_0 + B_1 L + \ldots + B_k L^k + \ldots$, where $L^k Y_t = Y_{t-k}$. Since $B(0) = I$, equation (1) is an unrestricted VAR representation of the joint dynamic behavior of the vector Y_t. In Sims's (1980a) original notation, the vector $e_t = (e_{yt},\ e_{xt})'$ would carry the meaning of "surprises" or innovations in Δy_t and Δx_t respectively.

In its simplest interpretation, the reduced form in (1) is a model that describes how the historical data contained in Y_t was generated by some random mechanism. As such, few would question its usefulness as a forecasting tool. However, in the analysis of the variables' dynamic responses to the various innovations, the implications of the unrestricted VAR are not unambiguous. Specifically, let us rewrite (1) as a moving average representation,

$$Y_t = B(L)^{-1} e_t = C(L) e_t, \tag{2}$$

where $C(L)$ is defined to be equal to $B(L)^{-1}$, with $C(L) = C_0 + C_1 L + \ldots + C_K L^K + \ldots$, and $C_0 = C(0) = B(0)^{-1} = I$. To obtain the comparative dynamic responses of Δy_t and Δx_t, Sims (1980a) first suggested orthogonalizing the vector of innovations e_t by defining $f_t = A e_t$, such that A is a lower triangular matrix with 1s on its diagonal and f_t has a normalized diagonal covariance

[1] Note that the details of the estimation procedure described in this article apply directly to the work of King and Watson (1997).

matrix. This particular transformation is known as a Choleski factorization and the newly defined innovations, $f_t = (f_{yt}, f_{xt})'$, have unit variance and are orthogonal. Equation (2) can therefore also be expressed as

$$Y_t = C(L)A^{-1}Ae_t = D(L)f_t, \tag{3}$$

with $D(L) = C_0A^{-1} + C_1A^{-1}L + \ldots + C_kA^{-1}L^k + \ldots$. Responses to innovations at different horizons, also known as impulse responses, are then given by

$$E_t\frac{\partial Y_{t+k}}{\partial f_t} = C_kA^{-1}, \text{ for } k = 0, 1, \ldots. \tag{4}$$

The advantage of computing dynamic responses in this way is that the innovations f_t are uncorrelated. Therefore it is very simple to compute the variances associated with any linear combinations involving them. Note that

$$E_{t+1}Y_{t+k} - E_tY_{t+k} = C_kA^{-1}f_t, \tag{5}$$

so that the j^{th} row of C_kA^{-1} gives the marginal effect of f_t on the j^{th} variable's k step-ahead forecast error. Since the f_t's are uncorrelated with unit variance, squaring the elements of C_kA^{-1} leads to contributions of the elements of f_t to the variance of the k step-ahead forecast error. This latter process is known as variance decomposition and describes the degree to which a particular innovation contributes to observed fluctuations in Y_t. Note that the variance decomposition of the contemporaneous forecast error is given by the squared elements of $C_0A^{-1} = A^{-1}$. More importantly, since A is a lower triangular matrix, A^{-1} is also lower triangular. This implies that the innovation in the first equation, f_{yt}, explains 100 percent of the variance in the contemporaneous forecast error of Δy_t. But this is precisely an identifying restriction on the dynamic behavior of Y_t. In a larger system, the variance of the contemporaneous forecast error in the j^{th} variable would be entirely accounted for by the first j innovations in a recursive fashion. Each of these restrictions would then implicitly constitute prior identifying restrictions. In this sense, the ordering of variables in a reduced-form VAR is of crucial significance.

This last point was made, perhaps most vigorously, in Cooley and Leroy (1985): "if the models (i.e., VARs) are interpreted as non-structural, we view the conclusions as unsupportable, being structural in nature. If the models are interpreted as structural, on the other hand, the restrictions on error distributions adopted in atheoretical macroeconometrics are not arbitrary renormalizations, but prior identifying restrictions." On a related note, Bernanke (1986) also writes that the standard Choleski decomposition, while "sometimes treated as neutral . . . in fact embodies strong assumptions about the underlying economic structure." Following these criticisms, several authors, including Blanchard and Watson (1984), Sims (1986), Bernanke (1986), and Blanchard and Quah (1989), addressed the issue of identification explicitly. The error terms in these latter models were given structural interpretations and the results no longer had to

depend on an arbitrary orthogonalization. However, this latter methodology possesses its own problems, both in terms of the validity of the estimation procedure and the interpretation of the results. This is the subject to which we now turn our attention.

2. INTRODUCTION TO THE MECHANICS OF STRUCTURAL VARs

The reduced form in equation (1) could simply be thought of as a way to summarize the full data set Y_t. In contrast, suppose that a theoretical model tells us that y_t actually evolves according to a specific stochastic process,

$$\Delta y_t = \Theta_{ya}(L)\epsilon_{at} + \Theta_{yb}(L)\epsilon_{bt} + (1-L)\Phi_{ya}(L)\epsilon_{at} + (1-L)\Phi_{yb}(L)\epsilon_{bt}, \qquad (6)$$

where ϵ_{at} and ϵ_{bt} now possess well-defined structural interpretations. Thus, y_t might represent national output, while ϵ_{at} and ϵ_{bt} might denote shocks to technology and labor supply respectively. This specification for y_t is quite general in that it allows shocks to have both permanent and temporary effects. The polynomial in the lag operator $\Phi(L)$ captures temporary deviations in y_t, while the polynomial $\Theta(L)$ keeps track of permanent changes in its steady-state level. Similarly, suppose that x_t follows a process that can be described by

$$\Delta x_t = \Theta_{xa}(L)\epsilon_{at} + \Theta_{xb}(L)\epsilon_{bt} + (1-L)\Phi_{xa}(L)\epsilon_{at} + (1-L)\Phi_{xb}(L)\epsilon_{bt}. \qquad (7)$$

With this specification in hand, it is possible to summarize the system as

$$Y_t = S(L)\epsilon_t, \qquad (8)$$

where Y_t is defined as in the previous section, $\epsilon_t = (\epsilon_{at}, \epsilon_{bt})'$, and

$$S(L) = \begin{bmatrix} \Theta_{ya}(L) + (1-L)\Phi_{ya}(L) & \Theta_{yb}(L) + (1-L)\Phi_{yb}(L) \\ \Theta_{xa}(L) + (1-L)\Phi_{xa}(L) & \Theta_{xb}(L) + (1-L)\Phi_{xb}(L) \end{bmatrix}. \qquad (9)$$

Equation (8) therefore denotes the structural moving average representation of the variables y_t and x_t, as a function of the exogenous innovations ϵ_{at} and ϵ_{bt}. Let us assume that $S(L)$ is invertible so that equation (8) can also be expressed in autoregressive form:

$$T(L)Y_t = S(L)^{-1}Y_t = \epsilon_t, \qquad (10)$$

that is,

$$T(L)\begin{bmatrix} \Delta y_t \\ \Delta x_t \end{bmatrix} = \begin{bmatrix} \epsilon_{at} \\ \epsilon_{bt} \end{bmatrix}, \text{ with } T(0) = S(0)^{-1} \neq I. \qquad (11)$$

Since the two exogenous processes that govern the behavior of y_t and x_t in (6) and (7) are assumed stationary, we also assume that the roots of the polynomial matrix $|T(z)|$ lie outside the unit circle. At this stage it is not possible to disentangle the structural effects of ϵ_{at} and ϵ_{bt} in equation (11). Put another

way, we cannot currently identify the structural error terms ϵ_{at} and ϵ_{bt} with the residuals in the two equations implicit in (11). This is a well-known problem that naturally leads us to the issue of identification.

Identification in Structural VARs

To get a handle on the problem of identification, observe the relationship between the reduced form in (1) and equation (11). Since $T(L)Y_t = T_0Y_t + T_1Y_{t-1} + \ldots$, it follows that $T_0^{-1}T(L)Y_t = Y_t + T_0^{-1}T_1Y_{t-1} + \ldots = T_0^{-1}\epsilon_t$. We then see that $T_0^{-1}T(L)Y_t$ is the reduced form, that is, $T_0^{-1}T(L) = B(L)$ so that

$$T(0)^{-1}T(L)Y_t = B(L)Y_t = e_t = T(0)^{-1}\epsilon_t. \tag{12}$$

Hence, if $\Sigma = \text{cov}\ (\epsilon_t)$ and $\Omega = \text{cov}\ (e_t)$, the following relation also holds:

$$T(0)^{-1}\Sigma T(0)^{-1\prime} = \Omega. \tag{13}$$

Since Ω can be estimated from the reduced form, the problem of identification relates to the conditions under which the structural parameters in $T(0)^{-1}\Sigma T(0)^{-1\prime}$ can be recovered from Ω. Equation (13) potentially establishes a set of three equations in seven unknowns. Specifically, the unknowns consist of four parameters in $T(0)$ and two variances and one covariance term in Σ. The SVAR literature typically reduces the size of this problem by making the following two assumptions. First, $T(0)$ is normalized to contain 1s on its diagonal. Second, Σ is diagonalized, which reflects the assumption that the structural disturbance terms are taken to be uncorrelated. This leaves us with four unknowns; therefore, one further restriction must be imposed for the structural form to be identified. This additional restriction will generally reflect the econometrician's beliefs and, as will be apparent below, will allow one to separate the effects of the two structural error terms.

As we have just pointed out, only one restriction needs to be imposed upon the dynamics of the system in (11) for the parameters to be identified. One possibility is to specify a priori one of the parameters in the contemporaneous matrix $T(0)$. Another popular approach, the one we focus on here, is to pre-specify a particular long-run relationship between the variables and therefore constrain the matrix of long-run multipliers $T(1)$. This approach is the one followed by Shapiro and Watson (1988), Blanchard and Quah (1989), King, Plosser, Stock, and Watson (1991), and Gali (1992, 1996) among others. To be concrete, define

$$T(1) = \begin{bmatrix} 1-\theta_{yy} & -\theta_{yx} \\ -\theta_{xy} & 1-\theta_{xx} \end{bmatrix} = \begin{bmatrix} \Theta_{ya}(1) & \Theta_{yb}(1) \\ \Theta_{xa}(1) & \Theta_{xb}(1) \end{bmatrix}^{-1} = S(1)^{-1}. \tag{14}$$

One way to achieve identification would be to impose the restriction that the exogenous process with innovation ϵ_{at} not affect the level of x_t in the long run.

That is, impose the restriction that

$$\Theta_{xa}(1) = 0. \tag{15}$$

Since inverses of block diagonal matrices are themselves block diagonal, setting $\Theta_{xa}(1) = 0$ is tantamount to setting $\theta_{xy} = 0$. It would then be possible to estimate all the remaining parameters in equation (6) and (7). This type of restriction, known as an exclusion restriction, is used for identification in the papers cited above. Note, however, that in theory there is no reason why identified parameters should be set to zero as opposed to any other value. All that is required is that the set of identified parameters be fixed in advance, whether zero or not. For example, if ϵ_{at} denotes a shock to technology and x_t represents labor supply, imposing $\Theta_{xa}(1) = 0$ would mean the structural model we have in mind implies that changes in technology do not affect labor supply in the long run. However, in a standard real business cycle model, the permanent effect of technology on labor supply depends on whether the income or the substitution effect dominates. This effect in turn depends on whether the elasticity of intertemporal substitution is greater or less than one. Therefore, there is no reason why exclusion restrictions should necessarily be used as an identification strategy.

The fact that $\Theta_{xa}(1)$, or alternatively θ_{xy}, does not have to be set to zero as a way to identify the model means that estimated parameters, and therefore estimated dynamic responses, can vary depending on the identification scheme adopted. This observation carries with it two potential problems. First, different identification schemes might lead to different comparative dynamic responses of the variables. Therefore, in using SVARs to establish stylized facts, some sensitivity analysis appears to be essential. Second, the estimation procedure may fail in a statistical sense for some values of θ_{xy} in the relevant parameter space. Before looking at each of these problems, however, we first need to explain SVAR estimation.

Structural VAR Estimation Procedure

The most popular way of imposing identifying restrictions as part of the estimation procedure in a SVAR is to take an instrumental variables (IV) approach, specifically two-stage least squares. In applying this approach to our bivariate system, we examine a simple case involving one lag. This will help in keeping matters tractable. Thus, the second equation in (11) can be written as

$$\Delta x_t = \beta_{xy0}\Delta y_t + \beta_{xy1}\Delta y_{t-1} + \beta_{xx1}\Delta x_{t-1} + \epsilon_{bt}. \tag{16}$$

To see how the long-run multipliers θ_{xx} and θ_{xy} in $T(1)$ implicitly enter in equation (16), observe that this equation can also be expressed as

$$\Delta x_t - \theta_{xy}\Delta y_t = \gamma_{xy0}\Delta^2 y_t + \theta_{xx}\Delta x_{t-1} + \epsilon_{bt}, \tag{17}$$

where $\Delta^2 y_t$ denotes the second difference in y_t, $\theta_{xx} = \beta_{xx1}$, $\gamma_{xy0} = -\beta_{xy1}$, and $\theta_{xy} = \beta_{xy0} + \beta_{xy1}$.[2] By setting a predetermined value for θ_{xy}, not necessarily zero, the parameters of equation (17) can then be estimated. Since $\Delta^2 y_t$ is correlated with ϵ_{bt}, ordinary least squares estimation is inappropriate, but two-stage least squares can be performed using the set $\mathbf{Z} = \{\Delta x_{t-1}, \Delta y_{t-1}\}$ as instruments. In a similar fashion, the equation for Δy_t can be written as

$$\Delta y_t = \beta_{yy1}\Delta y_{t-1} + \beta_{yx0}\Delta x_t + \beta_{yx1}\Delta x_{t-1} + \epsilon_{at}. \tag{18}$$

Equation (18) can be estimated using the same set of instruments as for (17) plus the estimated residual for ϵ_{bt}.[3] Recall that in order to achieve identification, the structural disturbances were assumed uncorrelated, thereby allowing the use of the estimated residual as an instrument. Furthermore, this residual is the only candidate instrument that remains. Additional lags of the endogenous variables, if relevant, should have been included in the original equations.

The key point to note at this stage is that since the left-hand side of equation (17) varies with θ_{xy}, the parameters as well as the error term in that equation are contingent upon the identification scheme. This raises a question as to the validity of the estimated residual from equation (17) as an instrument. Not only is zero correlation between the structural disturbances necessary, but a high correlation between the instrument and the variable it is instrumenting for is also essential. This point is emphasized by Nelson and Startz (1990). As we shall now see, because the time series behavior of the estimated residual in (17) varies with θ_{xy}, the validity of the estimation procedure in the subsequent equation will be implicitly tied to the choice of identifying restriction.

3. IDENTIFICATION FAILURE IN STRUCTURAL VARs

To gain insight into the problems that may arise in this framework, given the identification strategy adopted, let us rewrite equation (17) as follows:

$$\Delta x_t - \theta_{xy}\Delta y_t = \mathbf{X}\phi + \epsilon_{bt}, \tag{19}$$

where $\mathbf{X} = \{\Delta^2 y_t, \Delta x_{t-1}\}$ and $\phi = (\gamma_{xy0}, \theta_{xx})'$. Then, the two-stage least squares estimator $\widehat{\phi}$ is given by

$$\widehat{\phi} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'(\Delta x_t - \theta_{xy}\Delta y_t). \tag{20}$$

From equation (20), the parameter estimates in $\widehat{\phi}$ will change as θ_{xy} takes on different values. This is also true of the estimated residual, which we therefore

[2] As an intermediate step, equation (16) can also be expressed as $\Delta x_t = (\beta_{xy0} + \beta_{xy1} - \beta_{xy1})\Delta y_t + \beta_{xy1}\Delta y_{t-1} + \beta_{xx1}\Delta x_{t-1} + \epsilon_{bt}$.

[3] Observe that, analogously to (17), this equation can also be written as $\Delta y_t = \theta_{yy}\Delta y_{t-1} + \theta_{yx}\Delta x_t + \gamma_{yx0}\Delta^2 x_t + \epsilon_{at}$, where $\theta_{yy} = \beta_{yy1}$, $\gamma_{yx0} = -\beta_{yx1}$, and $\theta_{yx} = \beta_{yx0} + \beta_{yx1}$.

denote by $e_{bt}(\theta_{xy})$ to underscore its dependence on the adopted identification strategy. Since this estimated residual can be computed as

$$\begin{aligned} e_{bt}(\theta_{xy}) &= (\Delta x_t - \theta_{xy}\Delta y_t) - \mathbf{X}\widehat{\phi} \\ &= (\Delta x_t - \theta_{xy}\Delta y_t) - \mathbf{X}(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'(\Delta x_t - \theta_{xy}\Delta y_t), \end{aligned} \tag{21}$$

observe that $\mathbf{Z}'e_{bt}(\theta_{xy}) = e_{bt}(\theta_{xy})'\mathbf{Z} = 0 \ \forall\theta_{xy}$. This last condition summarizes what are sometimes called the normal equations. Now, the second equation to be estimated in (18) can also be expressed as

$$\Delta y_t = \mathbf{Z}\beta + \Delta x_t\beta_{yx0} + \epsilon_{at}, \tag{22}$$

where $\beta = (\beta_{yx1}, \beta_{yy1})'$, and Δx_t is the endogenous variable of interest. Since the relevant set of instruments for the estimation of equation (22) is given by $\{\mathbf{Z}, e_{bt}(\theta_{xy})\}$, it follows that the two-stage least squares estimator for β is given by

$$\begin{bmatrix} \widehat{\beta} \\ \widehat{\beta_{yx0}} \end{bmatrix} = \begin{bmatrix} \mathbf{Z}'\mathbf{Z} & \mathbf{Z}'\Delta x_t \\ e_{bt}(\theta_{xy})'\mathbf{Z} & e_{bt}(\theta_{xy})'\Delta x_t \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{Z}'\Delta y_t \\ e_{bt}(\theta_{xy})'\Delta y_t \end{bmatrix}. \tag{23}$$

This last expression can be thought of as a set of two equations in two unknowns, specifically,

$$\mathbf{Z}'\mathbf{Z}\widehat{\beta} + \mathbf{Z}'\Delta x_t\widehat{\beta_{yx0}} = \mathbf{Z}'\Delta y_t \tag{24}$$

and

$$e_{bt}(\theta_{xy})'\mathbf{Z}\widehat{\beta} + e_{bt}(\theta_{xy})'\Delta x_t\widehat{\beta_{yx0}} = e_{bt}(\theta_{xy})'\Delta y_t. \tag{25}$$

Therefore it follows that

$$\widehat{\beta_{yx0}} = [e_{bt}(\theta_{xy})'\mathbf{M}_z\Delta x_t]^{-1}[e_{bt}(\theta_{xy})'\mathbf{M}_z\Delta y_t], \tag{26}$$

where $\mathbf{M}_z$ is the projection matrix $\mathbf{I} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$. But we have just seen that $e_{bt}(\theta_{xy})'\mathbf{Z} = 0 \ \forall\theta_{xy}$, hence equation (26) simplifies to

$$\widehat{\beta_{yx0}} = [e_{bt}(\theta_{xy})'\Delta x_t]^{-1}[e_{bt}(\theta_{xy})'\Delta y_t]. \tag{27}$$

In other words, the two-stage least squares estimator for β_{yx0}, and hence the long-run multiplier θ_{yx}, depends on two key elements: the correlations of the estimated residual from the previous equation, equation (19), with both Δx_t and Δy_t. This is because each equation in a SVAR possesses many regressors in common. Since the "extra" instrument $e_{bt}(\theta_{xy})$ in the second equation is the residual from the first equation, it is by construction orthogonal to the other instruments in the second equation. It then follows that the two-stage least squares estimator for β_{yx0} depends only on the correlations of this residual with Δx_t and Δy_t as shown by (27). To see that certain identification schemes may be problematic, define θ^*_{xy} such that $e_{bt}(\theta^*_{xy})'\Delta x_t = 0$. Then, as long as $e_{bt}(\theta_{xy})'\Delta y_t$ remains finite, $\widehat{\beta_{yx0}}$ diverges when $\theta_{xy} \to \theta^*_{xy}$. In more standard IV settings, this result would not emerge. Residuals from other equations would not

generally be used as regressors, and hence parameter estimates would depend on more than one correlation.

To determine the exact value of the problematic identifying restriction, θ^*_{xy}, given the data under consideration, it suffices to take the transpose of equation (21), post-multiply the result by Δx_t, and set it to zero to yield

$$\theta^*_{xy} = \frac{\Delta x_t' \mathbf{W} \Delta x_t}{\Delta y_t' \mathbf{W} \Delta y_t}, \text{ where } \mathbf{W} = \mathbf{Z}(\mathbf{X}'\mathbf{Z})^{-1}\mathbf{X}' - \mathrm{I}. \tag{28}$$

To continue with our discussion, observe from equations (22) and (27) that

$$\widehat{\beta_{yx0}} - \beta_{yx0} = [e_{bt}(\theta_{xy})' \Delta x_t]^{-1} [e_{bt}(\theta_{xy})' \epsilon_{at}]. \tag{29}$$

Therefore a lower bound for the variance of the two-stage least squares estimator $\widehat{\beta_{yx0}}$ is given by

$$\text{var}\,(\widehat{\beta_{yx0}}) = \sigma^2_{\epsilon_a} [e_{bt}(\theta_{xy})' \Delta x_t]^{-1} [e_{bt}(\theta_{xy})' e_{bt}(\theta_{xy})] [e_{bt}(\theta_{xy})' \Delta x_t]^{-1\prime}, \tag{30}$$

where $\sigma^2_{\epsilon_a} = E(\epsilon^2_{at})$.[4] As $\theta_{xy} \to \theta^*_{xy}$, this variance diverges at the squared rate of that at which $\widehat{\beta_{yx0}}$ itself diverges. Taken together, equations (27) and (30) tell us that for identification strategies in a neighborhood of θ^*_{xy}, it is not possible to obtain a meaningful estimate of β_{xy0}. Both its estimator as well as associated confidence interval become arbitrarily large.

The above analysis has been numerical in nature in order to make clear the source of identification failure in SVAR estimation. One may wonder further, however, about the relationship between the distributional properties of $\widehat{\beta_{yx0}}$ and the identification restriction θ_{xy}. The questions of statistical inference and asymptotic distribution can be answered to some degree, it turns out, as a special case of the analysis carried out by Staiger and Stock (1993). Their analysis indicates that conventional asymptotic inference procedures are no longer valid when $e_{bt}(\theta_{xy})$ is weakly related to Δx_t in a regression of Δx_t on its instruments.[5]

Since residuals are recursively used as instruments in the estimation of SVARs, the "validity" of the estimation procedure implicitly depends on the nature of the identifying restrictions adopted. That is, the strength of the instruments is contingent upon the identification scheme. Some structural economic models may then be impossible to investigate empirically within the confines of a just-identified SVAR. In particular, as long as an identification strategy generates a small correlation between a recursively estimated residual and the variable it is meant to instrument for in the subsequent equation, coefficient estimates will lose their standard distributional properties.

[4] This is only a lower bound since $e_{bt}(\theta_{xy})$ is a generated regressor and therefore possesses some variation not accounted for in equation (30).

[5] See Appendix.

An Illustrative Example

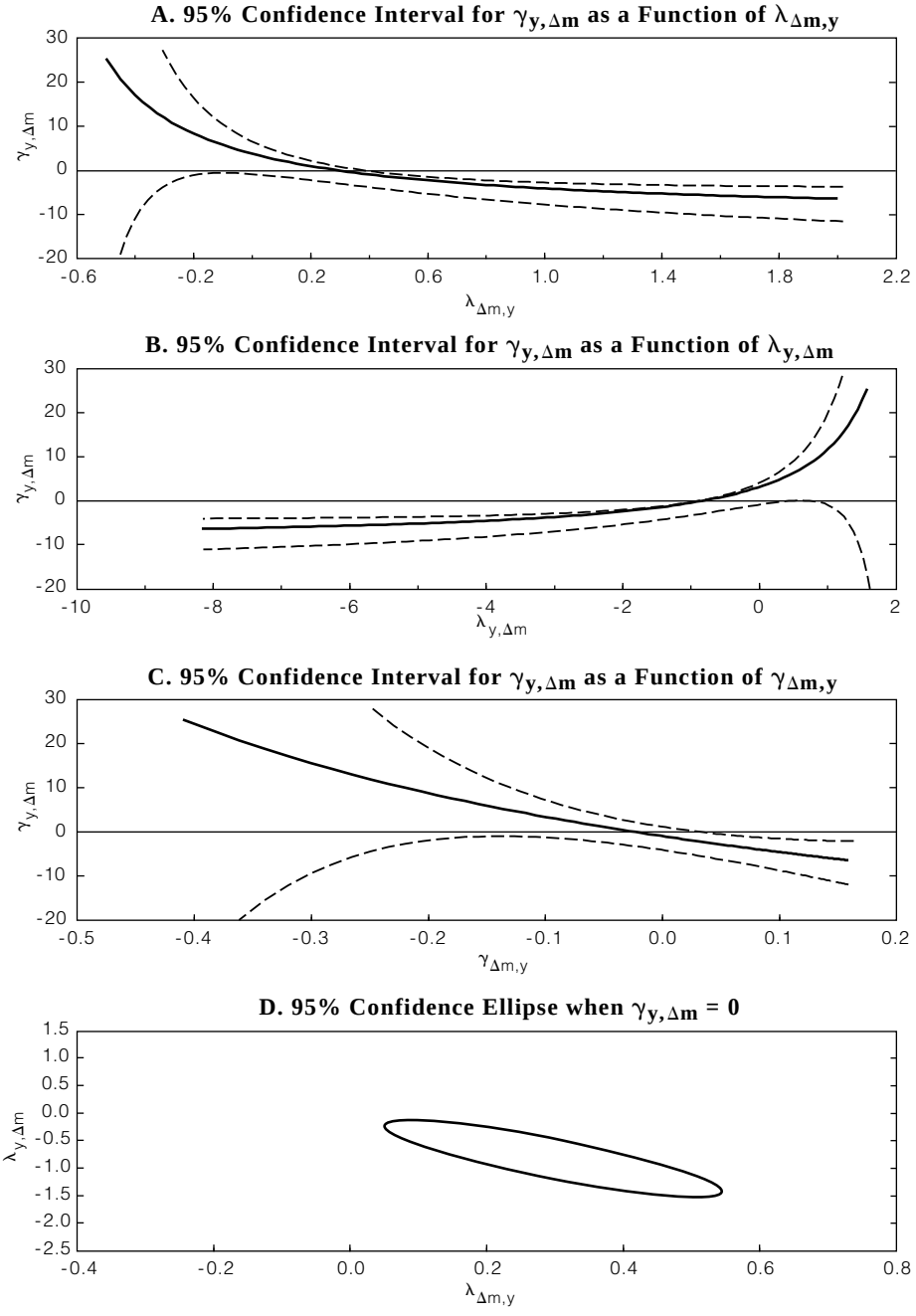
Although the analysis in this section has been carried out with a long-run identifying restriction in mind, the arguments above are also relevant in settings incorporating short-run identifying restrictions. As an example, consider a recent paper on long-run neutrality by King and Watson (1997). The authors estimate a bivariate system in output and money in order to test long-run money neutrality. In doing so, they recognize the importance of considering alternative identifying restrictions for robustness. A subset of their results are reproduced in Figure 1. In panel A of Figure 1, King and Watson (1997) report point estimates and confidence intervals for the hypothesis of long-run superneutrality when the short-run elasticity of money demand with respect to output is allowed to vary. Observe that as this value approaches -0.2, both the coefficient estimate for long-run superneutrality and its confidence intervals begin to blow up. In a similar fashion, panel C shows long-run superneutrality results under various assumptions with respect to the long-run response of money to exogenous permanent shifts in the level of output. Here, $\gamma_{\Delta m,y}$ corresponds to θ_{xy} so that in our notation, Δx_t is the money variable, while y_t is the output variable. As in the case where a short-run identifying restriction was considered, the estimate for long-run superneutrality and its associated confidence intervals start to diverge as $\gamma_{\Delta m,y}$ approaches -0.35. Thus, it should be clear that in looking for robustness across different identification schemes, one may be confronted with cases where the SVAR methodology cannot be meaningfully implemented.

At this stage, there remains at least one other obvious issue of interest. In our context, there may exist a plausible range of identifying restrictions in θ_{xy} for which the residual $e_{bt}(\theta_{xy})$ is, in fact, a proper instrument. If this were the case, one would naturally wonder whether comparative dynamic response estimates are sensitive to the particular identifying restriction imposed upon the system. The next section provides an example of interpretation ambiguities associated with precisely this issue.

4. INTERPRETING STRUCTURAL VARs: TECHNOLOGY SHOCKS AND AGGREGATE EMPLOYMENT FLUCTUATIONS

One topic of considerable interest in macroeconomics is the relationship between technology shocks and aggregate fluctuations in employment. Real business cycle models typically predict that technological innovations raise the level of employment. This result reflects the increase in the marginal productivity of labor associated with the positive technology shock when labor supply is relatively less variable. In a recent paper, however, Gali (1996) suggests that this feature of real business cycle models does not hold empirically. By using a bivariate SVAR in labor productivity and employment, he is able to show

Figure 1 Money Growth and Output



that technology shocks appear to induce a persistent *decline* in employment. Furthermore, labor productivity *increases* temporarily in response to demand shocks.

To motivate the identification of the particular SVAR he uses, Gali (1996) suggests a stylized model whose key features are monopolistic competition, predetermined prices, and variable effort.[6] In such a framework, a positive technology shock enhances labor productivity while leaving aggregate demand unchanged due to sticky prices. Employment must therefore fall. In addition, a positive demand shock would be met by a higher level of "unobserved" effort as well as higher "measured" employment. Given a strong enough effort response, labor productivity would temporarily rise. Formally, the structure of Gali's (1996) model implies that employment evolves according to

$$\Delta h_t = \Theta_{h\eta}(L)\eta_t + \Theta_{h\xi}(L)\xi_t + (1-L)\Phi_{h\eta}(L)\eta_t + (1-L)\Phi_{h\xi}(L)\xi_t, \qquad (31)$$

where η_t and ξ_t denote money growth and technology shocks respectively. Here, money growth shocks are associated with the management of aggregate demand by the monetary authority and hence serve as a proxy for demand shocks. Since technology shocks induce a persistent decline in employment, we have $\Theta_{h\xi}(1) < 0$. Similarly, labor productivity is given by

$$\Delta q_t = \Theta_{q\eta}(L)\eta_t + \Theta_{q\xi}(L)\xi_t + (1-L)\Phi_{q\eta}(L)\eta_t + (1-L)\Phi_{q\xi}(L)\xi_t, \qquad (32)$$

with $\Theta_{q\eta}(0) + \Phi_{q\eta}(0) > 0$ to capture the contemporaneous positive effect of a demand shock on labor productivity. As in Section 2, this system of equations can be summarized as

$$T(L)Y_t = \epsilon_t, \qquad (33)$$

where $Y_t = (\Delta h_t, \Delta q_t)'$, $\epsilon_t = (\eta_t, \xi_t)'$, and

$$T(L) = \begin{bmatrix} \Theta_{h\eta}(L) + (1-L)\Phi_{h\eta}(L) & \Theta_{h\xi}(L) + (1-L)\Phi_{h\xi}(L) \\ \Theta_{q\eta}(L) + (1-L)\Phi_{q\eta}(L) & \Theta_{q\xi}(L) + (1-L)\Phi_{q\xi}(L) \end{bmatrix}^{-1}. \qquad (34)$$

The key identifying restriction that Gali (1996) imposes upon the dynamics of his system is that demand shocks do not have a permanent effect on labor productivity. In terms of our earlier notation, we have

$$T(1) = \begin{bmatrix} 1-\theta_{hh} & -\theta_{hq} \\ -\theta_{qh} & 1-\theta_{qq} \end{bmatrix} = \begin{bmatrix} \Theta_{h\eta}(1) & \Theta_{h\xi}(1) \\ \Theta_{q\eta}(1) & \Theta_{q\xi}(1) \end{bmatrix}^{-1}, \qquad (35)$$

with

$$\Theta_{q\eta}(1) = \theta_{qh} = 0. \qquad (36)$$

[6] For the details of the model, refer to Gali (1996).

Figure 2 plots impulse response functions for the bivariate SVAR we have just described. The data comprise the log of hours worked in the nonfarm business sector as well as gross domestic product (in 1987 dollars), less gross domestic product in the farm sector. The log of productivity was hence computed as the log of gross domestic product, less the log of hours worked. Four lags were used in estimation and the sample period covers 1949:1 to 1992:4. As in Gali (1996), observe that the structural response of employment to a positive technology shock is negative, both in the short and long run. Furthermore, this is true even within a 90 percent confidence interval.[7] Note also that the contemporaneous response of productivity to a demand shock is positive and, by construction, eventually vanishes. Of course, since we have used data that is very similar to that used in the original study, these results are hardly surprising. However, Gali (1996) argues that since these estimates seem to hold for the majority of G7 countries, the impact "of technology shocks yields a picture which is hard to reconcile with the prediction of (real business cycle) models." This statement makes it clear that, among other results, the persistent employment decline in response to a technology shock is implicitly interpreted as a stylized fact. As we know, however, Gali's (1996) estimates derive from his choice of identification scheme; deviations from that scheme must be considered in order to decide what constitutes a stylized fact.

Alternative Identification Strategies

There are several different ways to think about Gali's (1996) initial SVAR set-up. First, supposing that aggregate demand shocks account for more than just money growth shocks, demand shocks may have a permanent impact on productivity. For instance, a permanent increase in taxes in a real business cycle model would yield an increase in the steady-state ratio of employment to capital. Given a standard production function with constant returns to scale, this increase in the ratio of labor to capital would necessarily be accompanied by a fall in labor productivity. This would invalidate the restriction that $\Theta_{q\eta}(1) = \theta_{qh} = 0$. Moreover, since θ_{qh} represents the long-run elasticity of productivity with respect to employment, it might not be unreasonable to expect that $\theta_{qh} < 0$. Figure 3 shows the impulse response functions that result in Gali's (1996) framework when θ_{qh} is set to -0.5. Under this alternative identification strategy, the response of employment to a technology shock is no longer negative. In fact, both the short- and long-run responses of employment are now positive. By comparing Figures 2b and 3b, observe that this latter result seems to hold even when standard errors are taken into account. That is,

[7] To construct the standard error bands, Monte Carlo simulations were done using draws from the normal distribution for each of the two structural innovations. One thousand Monte Carlo draws were carried out in each case.

Figure 2 Identification Assumption: Demand Shocks Have No Long-Run Impact on Productivity

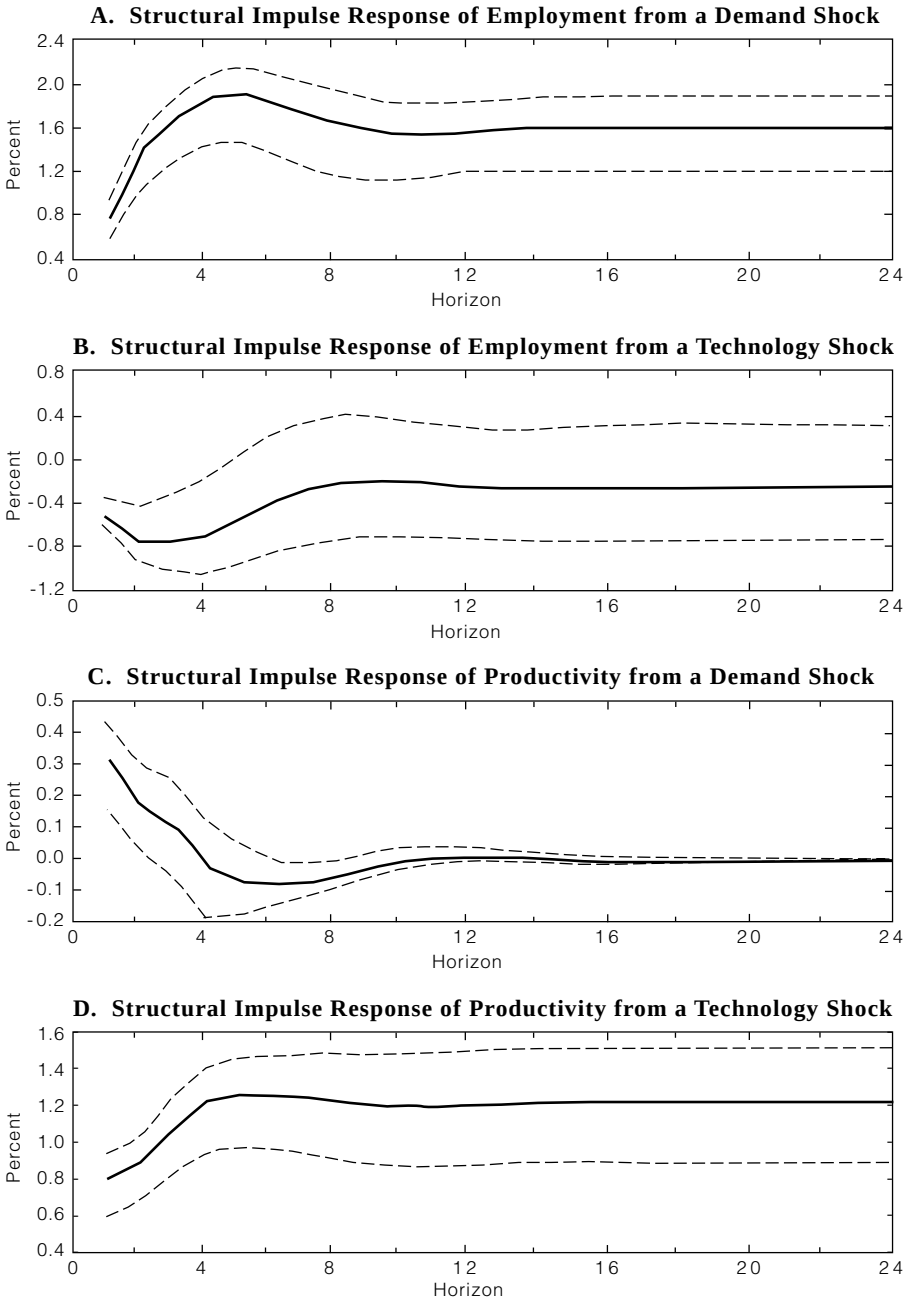
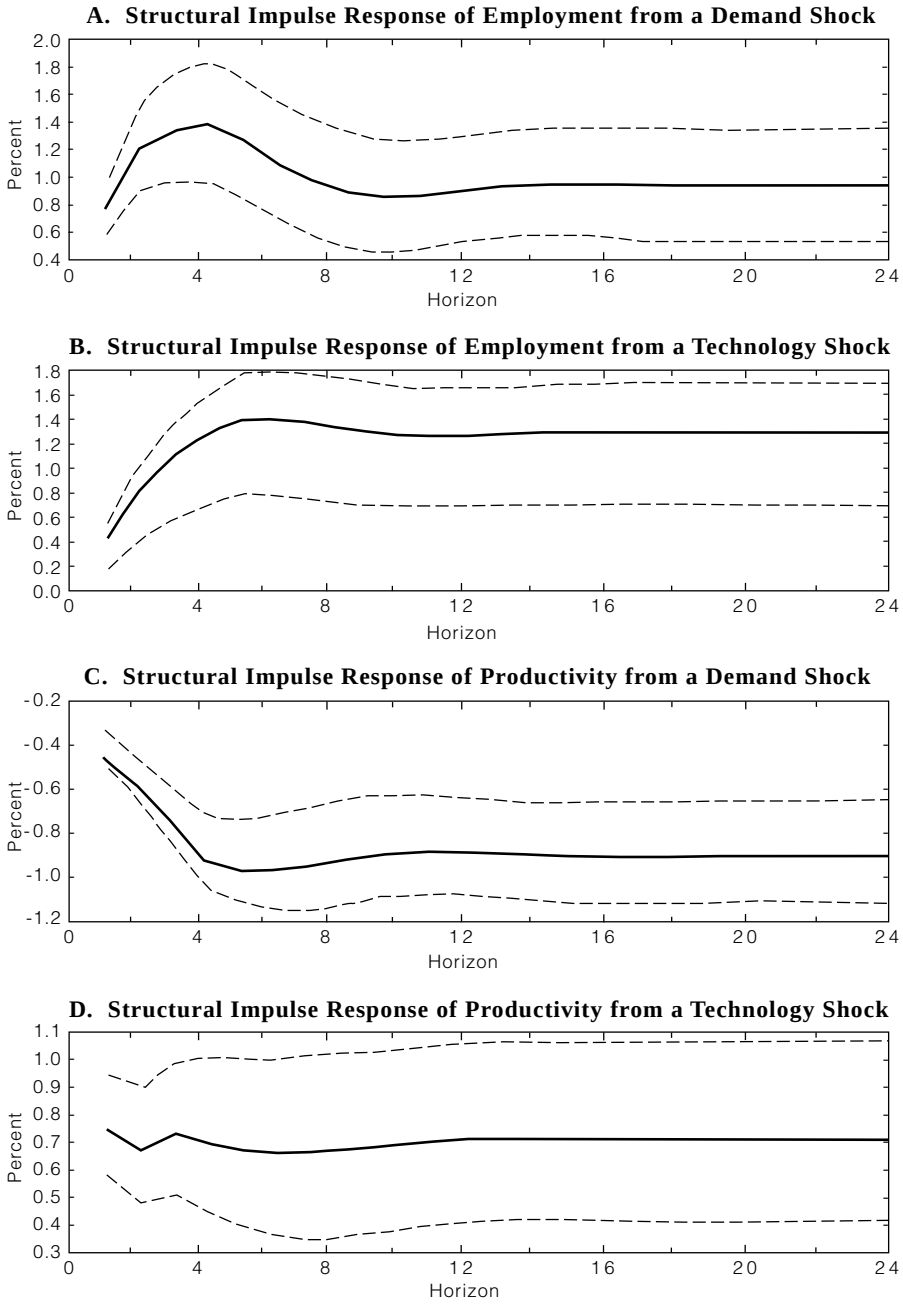


Figure 3 Identification Assumption: Demand Shocks Have a Negative Long-Run Impact on Productivity



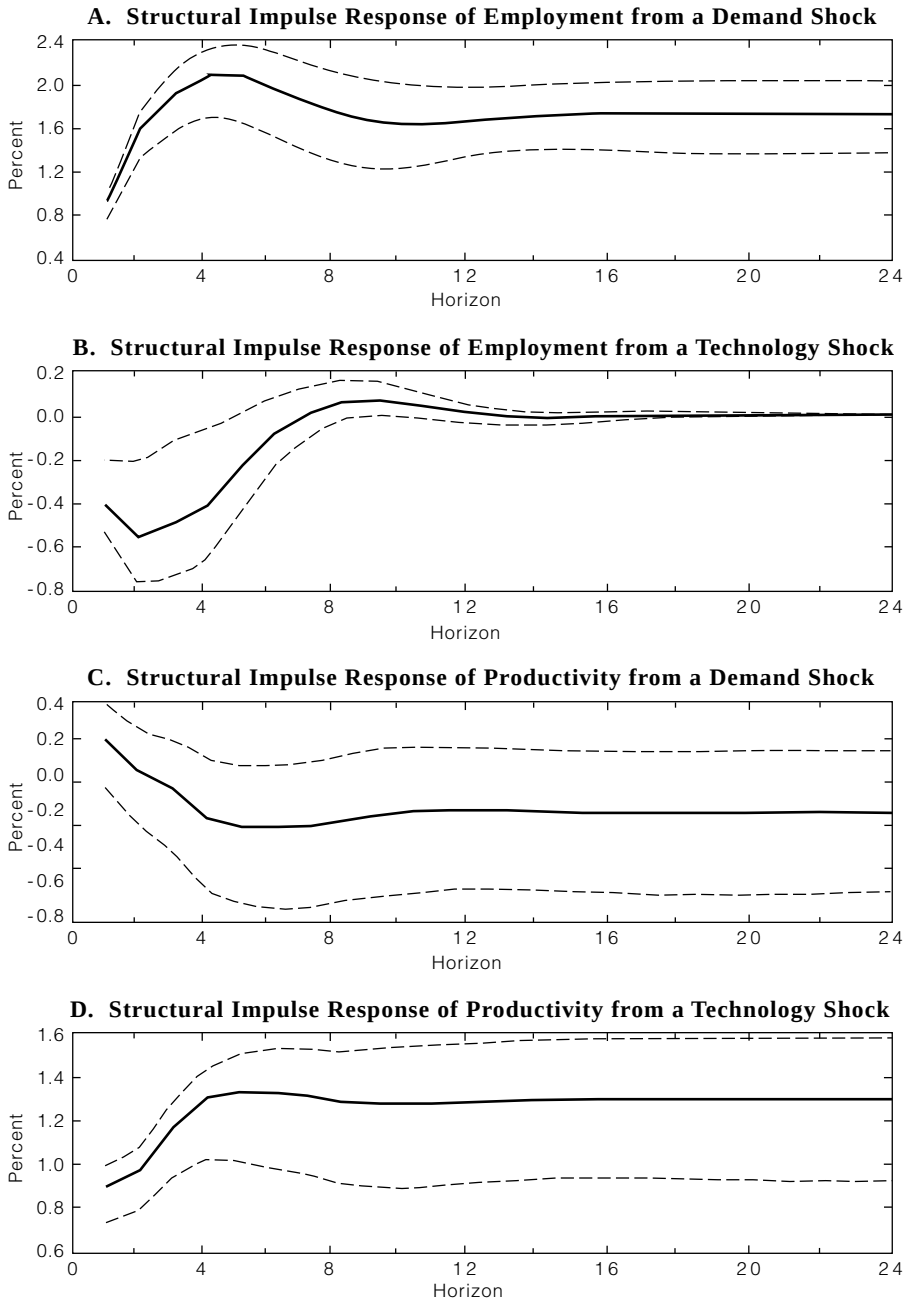
there is little overlap of the corresponding confidence intervals. Moreover, the contemporaneous effect of a demand shock on productivity is no longer positive but negative as shown in panel C. Viewed in this light, the dynamic response estimates initially reported in Gali (1996) may appear somewhat fragile. In particular, his contention that the data does not coincide with the predictions of real business cycle models does not necessarily hold.

In Figure 4, we show the results obtained when Gali's (1996) SVAR is identified using yet a third alternative. In this case, we require that technology shocks not have a long-run impact on employment. In terms of equation (35), this implies that $\Theta_{h\xi}(1) = \theta_{hq} = 0$. This identifying restriction is used by Shapiro and Watson (1988). It also emerges as a steady-state result in a real business cycle model when utility is logarithmic in consumption and leisure. Under this parameterization for utility, the income and substitution effects resulting from a positive technology shock cancel out, leaving labor supply unchanged in the steady state. (See King, Plosser, and Rebelo [1988].) Note in panel C of Figure 4 that under this third alternative, the long-run response of productivity to a demand shock is negative, which provides further evidence against Gali's (1996) initial identifying restriction. As already noted, this result is also consistent with a permanent increase in taxes in a real business cycle framework. Put another way, when one identifies Gali's (1996) bivariate system in a way that is consistent with the steady state generated by a standard real business cycle model, the empirical findings generated by the SVAR are consistent with the predictions of that model.

Of course, that is not to say that real business cycle models represent a more compelling framework when gauged against the data. The empirical results reported by Gali (1996) are themselves consistent with the theoretical model he uses to identify his SVAR. It is simply that in this case, what can be read from the data can vary sharply with one's prior beliefs concerning the theoretical nature of the data-generating mechanism.

While we have just shown that some of the key results in Gali (1996) are sensitive to the way one thinks about the long-run impact of various demand or supply shocks, this is not always the case. Observe that the structural impulse response of employment to a demand shock is similar in both direction and magnitude across Figures 2, 3, and 4. This is also true for the structural impulse response of productivity to a technology shock. Since these latter results emerge across estimated systems, that is, across systems with varying identifying restrictions, they may be reasonably considered stylized facts.

Figure 4 Identification Assumption: Technology Shocks Have No Long-Run Impact on Employment



5. SUMMARY AND CONCLUSIONS

We have investigated the extent to which identification issues can matter when using SVARs to characterize data. Although the main focus was on the estimation of bivariate systems, it should be clear that most of the above analysis applies to larger systems as well.

At a purely mechanical level, the source of the problem lies with the recursive use of an estimated residual as an instrument. The assumption made in SVAR estimation that the structural disturbances be uncorrelated is not sufficient to guarantee a proper estimation procedure. One must also pay attention to the degree of correlation between the estimated residual and the endogenous variable it is meant to be instrumenting for. This observation has long been made for simultaneous equations systems; and in this sense, it is important not to lose sight of the fact that SVARs are in effect a set of simultaneous equations.

At another level, we have also seen that even when the residual from a previously estimated equation is a valid instrument, SVARs can yield ambiguous results. This is the case even when confidence intervals are taken into account as in the bivariate example in hours and productivity. In that case, it was unclear whether employment responded positively or negatively, both in the short and long run, in response to a technology shock. Therefore, there may be a sense in which SVARs can fail in a way that is reminiscent of the Cooley and Leroy (1985) critique. In reduced-form VARs, different results emerge when alternative methods of orthogonalization of the error terms are adopted. In structural VARs, the results can now be directly contingent upon specific identifying restrictions. In effect, these are two facets of the same problem.

We have also seen in our example that certain results may be relatively robust with respect to the particular identification strategy of interest. For example, the response of productivity to a technology shock was estimated to be positive in both the short and long run across varying systems. Thus, two conclusions ultimately emerge from this investigation. First, special emphasis should be given to the derivation of identifying restrictions. The proper use of SVARs is contingent upon such restrictions and the case of identification failure cannot be ruled out a priori. Second, sensitivity analysis can be quite helpful in gaining a sense of the range of dynamics consistent with a given set of data. Assessing such a range seems an essential step in establishing stylized facts.

APPENDIX

This appendix derives the asymptotic distribution of $\widehat{\beta_{yx0}}$ in the text. This derivation is based on Staiger and Stock (1993). In the estimation of equation (22), suppose that the relationship that ties Δx_t to its instruments can be described as

$$\Delta x_t = \mathbf{Z}\alpha + e_{bt}(\theta_{xy})\alpha_{xe} + \nu_t, \tag{A1}$$

where ν_t is uncorrelated with ϵ_{at}. Furthermore, let us consider the set of identifying restrictions $\Pi_{\theta_{xy}}$ for which $\alpha_{xe} = N^{-1/2}g(\theta_{xy})$, where N is the sample size of our dataset and $g(\theta_{xy})$: $\Pi_{\theta_{xy}} \rightarrow \Re$. In other words, $\Pi_{\theta_{xy}}$ denotes a set of identifying restrictions for which the instrument $e_{bt}(\theta_{xy})$ is only weakly related to the endogenous variable Δx_t in the local to zero sense; the coefficient α_{xe} goes to zero as the sample size itself becomes arbitrarily large. To proceed with the argument, rewrite equation (29) as

$$\begin{aligned}\widehat{\beta_{yx0}} - \beta_{yx0} = \\ &[(N^{-1/2}\Delta x_t' e_{bt}(\theta_{xy}))(N^{-1}e_{bt}(\theta_{xy})'e_{bt}(\theta_{xy}))(N^{-1/2}e_{bt}(\theta_{xy})'\Delta x_t)]^{-1} \\ &[(N^{-1/2}\Delta x_t' e_{bt}(\theta_{xy}))(N^{-1}e_{bt}(\theta_{xy})'e_{bt}(\theta_{xy}))(N^{-1/2}e_{bt}(\theta_{xy})'\epsilon_{at})].\end{aligned} \tag{A2}$$

Given the assumptions embodied in (A1), it follows that

$$\begin{aligned}N^{-1/2}\Delta x_t' e_{bt}(\theta_{xy}) &= N^{-1/2}[\alpha' Z' + \alpha_{xe}e_{bt}(\theta_{xy}) + \nu_t']e_{bt}(\theta_{xy}) \\ &= N^{-1}e_{bt}(\theta_{xy})'e_{bt}(\theta_{xy})g(\theta_{xy}) + N^{-1/2}\nu_t' e_{bt}(\theta_{xy}).\end{aligned} \tag{A3}$$

Under suitable conditions, the first term in the above equation will converge to some constant almost surely as the sample size becomes large. The second term, on the other hand, will converge asymptotically to a normal distribution by the Central Limit Theorem. Therefore, although the coefficient on the relevant instrument, $e_{bt}(\theta_{xy})$, in the first-stage equation converges to zero, if the rate of convergence is slow enough, the right-hand side of equation (A2) will not diverge asymptotically. Nevertheless, in this case, the two-stage least squares estimator $\widehat{\beta_{yx0}}$ is asymptotically distributed as a ratio of quadratic forms in two jointly distributed normal variables. Hence, for identification strategies that belong to the set $\Pi_{\theta_{xy}}$, conventional asymptotic inference procedures will fail. In fact, in the so-called leading case where $g(\theta_{xy}) = 0$, Phillips (1989), Hillier (1985), and Staiger and Stock (1993) point out that $\widehat{\beta_{yx0}}$ asymptotically possesses a t distribution.

We now provide a sketch of the basic arguments. To this end, we assume that the following moment conditions are satisfied. The notation "$\rightarrow^p$" and "$\Rightarrow$" denote convergence in probability and convergence in distribution respectively.

(a) $(N^{-1}\mathrm{X}'\mathrm{X},\ N^{-1}\mathrm{Z}'\mathrm{X},\ N^{-1}\mathrm{X}'\Delta x_t,\ N^{-1}\mathrm{Z}'\Delta x_t,\ N^{-1}\mathrm{X}'\Delta y_t,\ N^{-1}\mathrm{Z}'\Delta y_t) \to^p$
$(\Sigma_{XX},\ \Sigma_{ZX},\ \Sigma_{X\Delta x_t},\ \Sigma_{Z\Delta x_t},\ \Sigma_{X\Delta y_t},\ \Sigma_{Z\Delta y_t})$

(b) $(N^{-1}\Delta x_t'\Delta x_t,\ N^{-1}\Delta y_t'\Delta y_t,\ N^{-1}\Delta x_t'\Delta y_t) \to^p (\Sigma_{\Delta x_t\Delta x_t},\ \Sigma_{\Delta y_t\Delta y_t},\ \Sigma_{\Delta x_t\Delta y_t})$

(c) $(N^{-1/2}\nu_t'\Delta x_t,\ N^{-1/2}\nu_t'\Delta y_t,\ N^{-1/2}\nu_t'\mathrm{X},\ N^{-1/2}\epsilon_{at}'\Delta x_t,\ N^{-1/2}\epsilon_{at}'\Delta y_t,$
$N^{-1/2}\epsilon_{at}'\mathrm{X}) \Rightarrow (\Psi_{\nu_t\Delta x_t},\ \Psi_{\nu_t\Delta y_t},\ \Psi_{\nu_t X},\ \Psi_{\epsilon_a\Delta x_t},\ \Psi_{\epsilon_a\Delta y_t},\ \Psi_{\epsilon_a X})$.

Note two particular points embodied in assumptions (a) through (c). First, assumptions (a) and (b) would naturally hold under standard conditions governing stationarity and ergodicity of the variables in the reduced form. Second, since these are primary assumptions, they do not depend on the identifying restriction θ_{xy}. It now remains to specify the asymptotic properties of three terms in (A2) and (A3), namely $N^{-1}e_{bt}(\theta_{xy})'e_{bt}(\theta_{xy})$, $N^{-1/2}\nu_t'e_{bt}(\theta_{xy})$, and $N^{-1/2}e_{bt}(\theta_{xy})'\epsilon_{at}$, to determine the asymptotic behavior of $\widehat{\beta_{yx0}(\theta_{xy})} - \beta_{yx0}(\theta_{xy})$ when $\theta_{xy} \in \Pi_{\theta_{xy}}$. Let us then examine each of these terms in turn.

Recall from equation (21) that

$$e_{bt}(\theta_{xy}) = (\Delta x_t - \theta_{xy}\Delta y_t) - \mathrm{X}(\mathrm{Z}'\mathrm{X})^{-1}\mathrm{Z}'(\Delta x_t - \theta_{xy}\Delta y_t).$$

It follows that $N^{-1}e_{bt}(\theta_{xy})'e_{bt}(\theta_{xy})$ is quadratic in θ_{xy}. Therefore, under assumptions (a) and (b), $N^{-1}e_{bt}(\theta_{xy})'e_{bt}(\theta_{xy}) \to^p \Sigma(\theta_{xy})$ uniformly, where $\Sigma(\theta_{xy})$ also depends on Σ_{XX}, Σ_{ZX}, etc. Next, consider $N^{-1/2}\nu_t'e_{bt}(\theta_{xy})$. We have

$$N^{-1/2}\nu_t'e_{bt}(\theta_{xy}) = N^{-1/2}[\nu_t'\Delta x_t - \theta_{xy}\nu_t'\Delta y_t - \nu_t'\mathrm{X}(\mathrm{Z}'\mathrm{X})^{-1}\mathrm{Z}'(\Delta x_t - \theta_{xy}\Delta y_t)],$$

which is linear in θ_{xy}. Therefore $N^{-1/2}\nu_t'e_{bt}(\theta_{xy}) \Rightarrow \Psi_{\nu_t}(\theta_{xy})$ uniformly, where $\Psi_{\nu_t}(\theta_{xy}) = \Psi_{\nu_t\Delta x_t} - \theta_{xy}\Psi_{\nu_t\Delta y_t} - \Psi_{\nu_t X}[\Sigma_{ZX}^{-1}\Sigma_{X\Delta x_t} - \theta_{xy}\Sigma_{ZX}^{-1}\Sigma_{Z\Delta y_t}]$. Finally, $N^{-1/2}e_{bt}(\theta_{xy})'\epsilon_{at}$ is given by

$$N^{-1/2}[\Delta x_t'\epsilon_{at} - \theta_{xy}\Delta y_t'\epsilon_{at} - (\Delta x_t' - \theta_{xy}\Delta y_t')\mathrm{Z}(\mathrm{X}'\mathrm{Z})^{-1}\mathrm{X}'\epsilon_{at}],$$

which is also linear in θ_{xy}. Hence, $N^{-1/2}e_{bt}(\theta_{xy})'\epsilon_{at} \Rightarrow \Psi_{\epsilon_{a_t}}(\theta_{xy})$ uniformly, where $\Psi_{\epsilon_{a_t}}(\theta_{xy}) = \Psi_{\epsilon_{a_t}\Delta x_t} - \theta_{xy}\Psi_{\epsilon_{a_t}\Delta y_t} - [\Sigma_{X\Delta x_t}'\Sigma_{ZX}^{-1\prime} - \theta_{xy}\Sigma_{X\Delta y_t}'\Sigma_{ZX}^{-1\prime}]\Psi_{\epsilon_{a_t}X}$. With these results in mind, it follows that $\widehat{\beta_{yx0}(\theta_{xy})}$ converges in distribution to

$$\beta_{yx0}(\theta_{xy}) +$$
$$[(g(\theta_{xy})\Sigma(\theta_{xy})^{1/2} + \Psi_{\nu_t}(\theta_{xy})\Sigma(\theta_{xy})^{-1/2})'(g(\theta_{xy})\Sigma(\theta_{xy})^{1/2} + \Psi_{\nu_t}(\theta_{xy})\Sigma(\theta_{xy})^{-1/2})]^{-1}$$
$$[(g(\theta_{xy})\Sigma(\theta_{xy})^{1/2} + \Psi_{\nu_t}(\theta_{xy})\Sigma(\theta_{xy})^{-1/2})'(\Sigma(\theta_{xy})^{-1/2}\Psi_{\epsilon_{a_t}}(\theta_{xy})].$$

This implies that for identification schemes in $\Pi_{\theta_{xy}}$, the two-stage least squares estimator is not only biased, it is asymptotically distributed as a ratio of quadratic forms in the jointly distributed normal random variables $\Psi_{\nu_t}(\theta_{xy})$ and $\Psi_{\epsilon_{a_t}}(\theta_{xy})$.

REFERENCES

Bernanke, Ben S. "Alternative Explanations of the Money-Income Correlation," *Carnegie-Rochester Conference Series on Public Policy*, vol. 25 (Autumn 1986), pp. 49–99.

Blanchard, Olivier J., and Danny Quah. "The Dynamic Effects of Aggregate Demand and Supply Disturbances," *American Economic Review*, vol. 79 (September 1989), pp. 655–73.

Blanchard, Olivier J., and Mark W. Watson. "Are Business Cycles All Alike?" in Robert J. Gordon, ed., *The American Business Cycle: Continuity and Change*, Chicago: University of Chicago Press, 1984.

Cooley, Thomas F., and Stephen F. Leroy. "Atheoretical Macroeconometrics: A Critique," *Journal of Monetary Economics*, vol. 16 (June 1985), pp. 283–308.

Gali, Jordi. "Technology, Employment, and the Business Cycle: Do Technology Shocks Explain Aggregate Fluctuations?" Mimeo, New York University, 1996.

_____. "How Well Does the IS-LM Model Fit Postwar U.S. Data?" *Quarterly Journal of Economics*, vol. 107 (May 1992), pp. 709–38.

Hillier, Grant H. "On the Joint and Marginal Densities of Instrumental Variables Estimators in a General Structural Equation," *Econometric Theory*, vol. 1 (April 1985), pp. 53–72.

King, Robert J., Charles I. Plosser, and Sergio T. Rebelo. "Production, Growth, and Business Cycles: II. New Directions," *Journal of Monetary Economics*, vol. 21 (March 1988), pp. 309–42.

King, Robert J., Charles I. Plosser, James H. Stock, and Mark W. Watson. "Stochastic Trends and Economic Fluctuations," *American Economic Review*, vol. 81 (September 1991), pp. 819–40.

King, Robert J., and Mark W. Watson. "Testing Long-Run Neutrality," Federal Reserve Bank of Richmond *Economic Quarterly*, vol. 83 (Summer 1997), pp. 69–101.

Nelson, Charles R., and Richard Startz. "The Distribution of the Instrumental Variables Estimator and its t-Ratio when the Instrument Is a Poor One," *Journal of Business*, vol. 63 (January 1990), pp. 125–40.

Phillips, Peter. C. "Partially Identified Models," *Econometric Theory*, vol. 5 (August 1989), pp. 181–240.

Shapiro, Mathew D., and Mark W. Watson. "Sources of Business Cycle Fluctuations," *NBER* Working Paper 1246, 1988.

Sims, Christopher. A. "Are Forecasting Models Usable for Policy Analysis?" Federal Reserve Bank of Minneapolis *Quarterly Review*, vol. 10 (Winter 1986), pp. 2–16.

_____. "Comparison of Interwar and Postwar Business Cycles: Monetarism Reconsidered," *American Economic Review*, vol. 70 (May 1980a), pp. 250–59.

_____. "Macroeconomics and Reality," *Econometrica*, vol. 48 (January 1980b), pp. 1– 47.

Staiger, Douglas, and James H. Stock. "Instrumental Variables Regressions with Weak Instruments." Mimeo, Kennedy School of Government, Harvard University, 1993.

Testing Long-Run Neutrality

Robert G. King and Mark W. Watson

Key classical macroeconomic hypotheses specify that permanent changes in nominal variables have no effect on real economic variables in the long run. The simplest "long-run neutrality" proposition specifies that a permanent change in the money stock has no long-run consequences for the level of real output. Other classical hypotheses specify that a permanent change in the rate of inflation has no long-run effect on unemployment (a vertical long-run Phillips curve) or real interest rates (the long-run Fisher relation). In this article we provide an econometric framework for studying these classical propositions and use the framework to investigate their relevance for the postwar U.S. experience.

Testing these propositions is a subtle matter. For example, Lucas (1972) and Sargent (1971) provide examples in which it is impossible to test long-run neutrality using reduced-form econometric methods. Their examples feature rational expectations together with short-run nonneutrality and exogenous variables that follow stationary processes so that the data generated by these models do not contain the sustained changes necessary to directly test long-run neutrality. In the context of these models, Lucas and Sargent argued that it was necessary to construct fully articulated behavioral models to test the neutrality propositions. McCallum (1984) extended these arguments and showed that low-frequency band spectral estimators calculated from reduced-form models were also subject to the Lucas-Sargent critique. While these arguments stand on firm logical ground, empirical analysis following the Lucas-Sargent prescriptions has not yet yielded convincing evidence on the neutrality propositions. This undoubtedly reflects a lack of consensus among macroeconomists on the appropriate behavioral model to use for the investigation.

The authors thank Marianne Baxter, Michael Dotsey, Robert Hetzel, Thomas Humphrey, Bennett McCallum, Yash Mehra, James Stock, and many seminar participants for useful comments and suggestions. This research was supported in part by National Science Foundation grants SES-89-10601, SES-91-22463, and SBR-9409629. The views expressed are those of the authors and do not necessarily reflect those of the Federal Reserve Bank of Richmond or the Federal Reserve System.

The specific critique offered by Lucas and Sargent depends critically on stationarity. In models in which nominal variables follow integrated variables processes, long-run neutrality can be defined and tested without complete knowledge of the behavioral model. Sargent (1971) makes this point clearly in his paper, and it is discussed in detail in Fisher and Seater (1993).[1] But, even when variables are integrated, long-run neutrality cannot be tested using a reduced-form model. Instead, what is required is the model's "final form," showing the dynamic response of the variables to underlying structural disturbances.[2]

Standard results from the econometric analysis of simultaneous equations show that the final form of a structural model is not econometrically identified, in general, because a set of a priori restrictions are necessary to identify the structural disturbances. Our objective in this article is to summarize the reduced-form information in the postwar U.S. data and relate it to the long-run neutrality propositions under alternative identifying restrictions. We do this by systematically investigating a wide range of a priori restrictions and asking which restrictions lead to rejections of long-run neutrality and which do not. For example, in our framework the estimated value of the long-run elasticity of output with respect to money depends critically on what is assumed about one of three other elasticities: (i) the impact elasticity of output with respect to money, (ii) the impact elasticity of money with respect to output, or (iii) the long-run elasticity of money with respect to output. We present neutrality test results for a wide range of values for these elasticities, using graphical methods.

Our procedure stands in stark contrast to the traditional method of exploring a small number of alternative identifying restrictions, and it has consequent costs and benefits. The key benefit is the extent of the information conveyed: researchers with strong views about plausible values of key parameters can learn about the result of a neutrality test appropriate for their beliefs; other researchers can learn about what range of parameter values result in particular conclusions about neutrality. The key cost is that the methods that we use are only practical in small models, and we demonstrate them here using

[1] Also see Geweke (1986), Stock and Watson (1988), King, Plosser, Stock, and Watson (1991), and Gali (1992).

[2] Throughout this article we use the traditional jargon of dynamic linear simultaneous equations. By "structural model" we mean a simultaneous equations model in which each endogenous variable is expressed as a function of the other endogenous variables, exogenous variables, lags of the variables, and disturbances that have structural interpretation. By "reduced-form model" we mean a set of regression equations in which each endogenous variable is expressed as a function of lagged dependent variables and exogenous variables. By "final-form model" we mean a set of equations in which the endogenous variables are expressed as a function of current and lagged values of shocks and exogenous variables in the model. For the standard textbook discussion of these terms, see Goldberger (1964), chapter 7.

bivariate models. This raises important questions about effects of potential omitted variables, and we discuss this issue below in the context of specific empirical models.

We organize our discussion as follows. In Section 1 below, we begin with the theoretical problem of testing for neutrality in economies that are consistent with the Lucas-Sargent conclusions. Our goal is to show the restrictions that long-run neutrality impose on the final-form model, and how these restrictions are related to the degree of integration of the variables. In Section 2, we discuss issues of econometric identification. Section 3 contains an empirical investigation of (i) the long-run neutrality of money, (ii) the long-run superneutrality of money, and (iii) the long-run Fisher relation. Even with an unlimited amount of data, the identification problems discussed above make it impossible to carry out a definitive test of the long-run propositions. Instead, we investigate the plausibility of the propositions across a wide range of observationally equivalent models. In Section 4 we investigate the long-run relation between inflation and the unemployment rate, i.e., the slope of the long-run Phillips curve. Here, the identification problem is more subtle than in the other examples. As we show, the estimated long-run relationship depends in an important way on whether the Phillips curve slope is calculated from a "supply" equation, as in Sargent (1976) for example, or from a "price" equation, as in Solow (1969) or Gordon (1970).

Previewing our empirical results, we find unambiguous evidence supporting the neutrality of money but more qualified support for the other propositions. Over a wide range of identifying assumptions, we find there is little evidence in the data against the hypothesis that money is neutral in the long run. Thus the finding that money is neutral in the long run is robust to a wide range of identifying assumptions. Conclusions about the other long-run neutrality propositions are not as unambiguous: these propositions are rejected for a range of identifying restrictions that we find arguably reasonable, but they are not rejected for others. Yet many general conclusions are robust. For example, the rejections of the long-run Fisher effect suggest that a one percentage point permanent increase in inflation leads to a smaller than one percentage point increase in nominal interest rates. Moreover, a wide range of identifying restrictions leads to very small estimates of the long-run effect of inflation on unemployment. On the other hand, the sign and magnitude of the estimated long-run effect of money growth on the level of output depends critically on the specific identifying restriction employed.

1. THE ROLE OF UNIT ROOTS IN TESTS FOR LONG-RUN NEUTRALITY

Early empirical researchers investigated long-run neutrality by examining the coefficients in the distributed lag:

$$y_t = \Sigma\alpha_j m_{t-j} + \text{error} = \alpha(L)m_t + \text{error}, \tag{1}$$

where y is logarithm of output, m is logarithm of the money supply, $\alpha(L) = \Sigma\alpha_j L^j$, and L is the lag operator.[3] If m_t is increased by one unit permanently, then (1) implies that y_t will eventually increase by the sum of the α_j coefficients. Hence, investigating the long-run multiplier, $\alpha(1) = \Sigma\alpha_j$, appears to be a reasonable procedure for investigating long-run neutrality. However, Lucas (1972) and Sargent (1971) demonstrated that in models with short-run nonneutrality and rational expectations, this approach can be very misguided.

The Lucas-Sargent critique can be exposited as follows. Consider a model consisting of an aggregate supply schedule (2a); a monetary equilibrium condition (2b); and a money supply rule (2c):

$$y_t = \theta(p_t - E_{t-1}p_t), \tag{2a}$$

$$p_t = m_t - \delta y_t, \text{ and} \tag{2b}$$

$$m_t = \rho m_{t-1} + \epsilon_t^m, \tag{2c}$$

where y_t is the logarithm of output; p_t is the logarithm of the price level; $E_{t-1}p_t$ is the expectation of p_t formed at $t-1$, m_t is the logarithm of the money stock, and ϵ_t^m is a mean-zero serially independent shock to money. The solution for output is

$$y_t = \pi(m_t - E_{t-1}m_t) = \pi(m_t - \rho m_{t-1}) = \pi(1 - \rho L)m_t = \alpha(L)m_t, \tag{3}$$

with $\pi = \theta/(1 + \delta\theta)$ and $\alpha(L) = \alpha_0 + \alpha_1 L = \pi(1 - \rho L)$.

As in Lucas (1973), the model is constructed so that only surprises in the money stock are nonneutral and these have temporary real effects. Permanent changes in money have no long-run effect on output. However, the reduced-form equation $y_t = \alpha(L)m_t$ suggests that a one-unit permanent increase in money will increase output by $\alpha_0 + \alpha_1 = \alpha(1) = \pi(1 - \rho)$. Moreover, as noted by McCallum (1984), the reduced form also implies that there is a long-run correlation between money and output, as measured by the spectral density matrix of the variables at frequency zero.

On this basis, Lucas (1972), Sargent (1971), and McCallum (1984) argue that a valid test of long-run neutrality can only be conducted by determining the structure of monetary policy (ρ) and its interaction with the short-run response to monetary shocks (π), which depends on the behavioral relations in the model (δ and θ). While this is easy enough to determine in this simple setting, it is much more difficult in richer dynamic models or in models with a more sophisticated specification of monetary policy.

[3] See Sargent (1971) for references to these early empirical analyses.

However, if $\rho = 1$, there is a straightforward test of the long-run neutrality proposition in this simple model. Adding and subtracting ρm_t from the right-hand side of (3) yields

$$y_t = \pi\rho\Delta m_t + \pi(1-\rho)m_t \tag{3'}$$

so that with $\rho = 1$ there is a zero effect of the level of money under the neutrality restriction. Hence, one can simply examine whether the coefficient on the level of money is zero when m_t is included in a bivariate regression that also involves Δm_t as a regressor.

With permanent variations in the money stock, the reduced form of this simple model has two key properties: (i) the coefficient on m_t corresponds to the experiment of permanently changing the level of the money stock; and (ii) the coefficient on Δm_t captures the short-run nonneutrality of monetary shocks. Equivalently, with $\rho = 1$, the neutrality hypothesis implies that in the specification $y_t = \Sigma\alpha_j m_{t-j}$, the neutrality restriction is $\alpha(1) = 0$, where $\alpha(1) = \Sigma\alpha_j$ is the sum of the distributed lag coefficients.

While the model in (2a) – (2c) is useful for expositing the Lucas-Sargent critique, it is far too simple to be used in empirical analysis. Standard macroeconomic models include several other important features: shocks other than ϵ_t^m are incorporated to capture other sources of fluctuations; the simple specification of an exogenous money supply in (2c) is discarded in favor of a specification that allows the money supply to respond to the endogenous variables in the model; and finally, the dynamics of the model are generalized through the incorporation of sticky prices, costs of adjusting output, information lags, etc. In these more general settings, it is still the case that long-run neutrality can sometimes be determined by examining the model's final form.

To see this, consider a macroeconomic model that is linear in both the observed variables and the structural shocks. Then, if the growth rates of both output and money are stationary, the model's final form can be written as

$$\Delta y_t = \mu_y + \theta_{y\eta}(L)\epsilon_t^\eta + \theta_{ym}(L)\epsilon_t^m \text{ and} \tag{4a}$$

$$\Delta m_t = \mu_m + \theta_{m\eta}(L)\epsilon_t^\eta + \theta_{mm}(L)\epsilon_t^m, \tag{4b}$$

where ϵ_t^η is vector of shocks, other than money, that affect output; $\theta_{mm}(L)\epsilon_t^m = \Sigma\theta_{mm,j}\epsilon_{t-j}^m$, and the other terms are similarly defined. Rich dynamics are incorporated in the model via the lag polynomials $\theta_{y\eta}(L)$, $\theta_{ym}(L)$, $\theta_{m\eta}(L)$, and $\theta_{mm}(L)$. These final-form lag polynomials will be functions of the model's behavioral parameters in a way that depends on the specifics of the model, but the particular functional relation need not concern us here.

The long-run neutrality tests that we conduct all involve the answer to the following question: does an unexpected and exogenous permanent change in the level of m lead to a permanent change in the level of y? If the answer is no, then we say that m is long-run neutral towards y. In equations (4a) and (4b), ϵ_t^m

are exogenous unexpected changes in money. The permanent effect of ϵ_t^m on future values of m is given by $\Sigma\theta_{mm,j}\epsilon_t^m = \theta_{mm}(1)\epsilon_t^m$. Similarly, the permanent effect of ϵ_t^m on future values of y is given by $\Sigma\theta_{ym,j}\epsilon_t^m = \theta_{ym}(1)\epsilon_t^m$. Thus, the long-run elasticity of output with respect to permanent exogenous changes in money is

$$\gamma_{ym} = \theta_{ym}(1)/\theta_{mm}(1). \tag{5}$$

Within this context, we say that the model exhibits long-run neutrality when $\gamma_{ym} = 0$. That is, the model exhibits long-run neutrality when the exogenous shocks that permanently alter money, ϵ_t^m, have no permanent effect on output.

In an earlier version of this article (King and Watson 1992) and in King and Watson (1994), we explored the relationship between the restriction $\gamma_{ym} = 0$ and the traditional notion of long-run neutrality using a dynamic linear rational expectations model with sluggish short-run price adjustment. We required that the model display theoretical neutrality, in that its real variables were invariant to proportionate changes in all nominal variables. We showed that this long-run neutrality requirement implied long-run neutrality in the sense investigated here. That is, unexpected permanent changes in m_t had no effect on y_t. Further, like the simple example presented in equations (2) and (3) above, the model also implied that long-run neutrality could be tested within a system like (4) if (and only if) the money stock is integrated of order one. Finally, in the theoretical model, long-run neutrality implied that $\gamma_{ym} = 0$.

In the context of equations (4a) – (4b), the long-run neutrality restriction $\gamma_{ym} = 0$ can only be investigated when money is integrated. If the money process does not contain a unit root, then there are no permanent changes in the level of m_t and $\theta_{mm}(1) = 0$. In this case, γ_{ym} in (5) is undefined, and the model's final form says nothing about long-run neutrality. This is the point of the Lucas-Sargent critique. The intuition underlying this result is simple: long-run neutrality asks whether a permanent change in money will lead to a permanent change in output. If permanent changes in money did not occur in the historical data (that is, money is stationary), then these data are uninformative about long-run neutrality. On the other hand, when the exogenous changes in money permanently alter the level of m, then $\theta_{mm}(1) \neq 0$, money has a unit root, γ_{ym} is well defined in (5), and the question of long-run neutrality can be answered from the final form of the model.

2. ECONOMETRIC ISSUES

In general, it is not possible to use data to determine the parameters of the final-form equations (4a) – (4b). Econometric identification problems must first be solved. We approach the identification problem in an unusual way. Rather

than "solve" it by imposing a single set of a priori restrictions, our empirical strategy is to investigate long-run neutrality for a large set of observationally equivalent models. Our hope is that this will provide researchers with a clearer sense of the robustness of any conclusions about long-run neutrality. Before presenting the empirical results, we review the issues of econometric identification that arise in the estimation of sets of equations like (4a) and (4b). This discussion motivates the set of observationally equivalent models analyzed in our empirical work.

To begin, assume that $(\epsilon_t^{\eta\prime}\epsilon_t^m)'$ is a vector of unobserved mean-zero serially independent random variables, so that (4a) – (4b) can be interpreted as a vector moving average model. The standard estimation strategy begins by inverting the moving average model to form a vector autoregressive model (VAR). The VAR, which is assumed to be finite order, is then analyzed as a dynamic linear simultaneous equations model.[4] We will work within this framework.

Estimation and inference in this framework requires two distinct sets of assumptions. The first set of assumptions is required to transform the vector moving average model into a VAR. The second set of assumptions is required to econometrically identify the parameters of the VAR. These sets of assumptions are intimately related: the moving average model can only be inverted if the VAR includes enough variables to reconstruct the structural shocks. In the context of (4a) – (4b), if $\epsilon_t = (\epsilon_t^{\eta\prime}\epsilon_t^m)'$ is an $n \times 1$ vector, then there must be at least n variables in the VAR. But, identification of an n-variable VAR requires $n \times (n-1)$ a priori restrictions, so that the necessary number of identifying restrictions increases with the square of the number of structural shocks.

In our empirical analysis we will assume that $n = 2$, so that only bivariate VARs are required. To us, this seems the natural starting point, and it has been employed by many other researchers in the study of the neutrality propositions discussed below. We also do this for tractability: when $n = 2$, only 2 identifying restrictions are necessary. This allows us to investigate thoroughly the set of observationally equivalent models. The cost of this simplification is that some of our results may be contaminated by omitted variables bias. We discuss this possibility more in the context of the empirical results.

To derive the set of observationally equivalent models, let $X_t = (\Delta y_t, \Delta m_t)'$, and stack (4a) – (4b) as

$$X_t = \Theta(L)\epsilon_t, \tag{6}$$

where $\epsilon_t = (\epsilon_t^{\eta}\epsilon_t^m)'$ is the 2×1 vector of structural disturbances. Assume that

[4] Standard references are Blanchard and Watson (1986), Bernanke (1986), and Sims (1986). See Watson (1994) for a survey.

$|\Theta(z)|$ has all of its zeros outside the unit circle, so that $\Theta(L)$ can be inverted to yield the VAR: [5]

$$\alpha(L)X_t = \epsilon_t, \tag{7}$$

where $\alpha(L) = \Sigma_{j=0}^{\infty}\alpha_j L^j$, with α_j a 2×2 matrix. Unstacking the Δy_t and Δm_t equations yields

$$\Delta y_t = \lambda_{ym}\Delta m_t + \sum_{j=1}^{p}\alpha_{j,yy}\Delta y_{t-j} + \sum_{j=1}^{p}\alpha_{j,ym}\Delta m_{t-j} + \epsilon_t^{\eta} \text{ and} \tag{8a}$$

$$\Delta m_t = \lambda_{my}\Delta y_t + \sum_{j=1}^{p}\alpha_{j,my}\Delta y_{t-j} + \sum_{j=1}^{p}\alpha_{j,mm}\Delta m_{t-j} + \epsilon_t^{m}, \tag{8b}$$

which is written under the assumption that the VAR in (7) is of order p.

Equation (7) or equivalently equations (8a) and (8b) are a set of dynamic simultaneous equations, and econometric identification can be studied in the usual way. Writing $\Sigma_\epsilon = E(\epsilon_t\epsilon_t')$, the reduced form of (7) is

$$X_t = \sum_{i=1}^{p}\Phi_i X_{t-i} + e_t, \tag{9}$$

where $\Phi_i = -\alpha_0^{-1}\alpha_i$ and $e_t = \alpha_0^{-1}\epsilon_t$. The matrices α_i and Σ_ϵ are determined by the set of equations

$$\alpha_0^{-1}\alpha_i = -\Phi_i, i = 1, \ldots, p \text{ and} \tag{10}$$

$$\alpha_0^{-1}\Sigma_\epsilon\alpha_0^{-1\prime} = \Sigma_e = E(e_t e_t'). \tag{11}$$

When there are no restrictions on coefficients on lags entering (9), equation (10) imposes no restrictions on α_0; it serves to determine α_i as a function of α_0 and Φ_i. Equation (11) determines both α_0 and Σ_ϵ as a function of Σ_e. Since Σ_e (a 2×2 symmetric matrix) has only three unique elements, only three unknown parameters in α_0 and Σ_ϵ can be identified. Equations (8a) and (8b) place 1s on the diagonal of α_0, but evidently only three of the remaining parameters $\text{var}(\epsilon_t^m), \text{var}(\epsilon_t^\eta), \text{cov}(\epsilon_t^m, \epsilon_t^\eta), \lambda_{my}$ and λ_{ym} can be identified. We follow the standard practice in structural VAR analysis and assume that the structural shocks are uncorrelated. Since λ_{my} and λ_{ym} are allowed to be nonzero, the assumption places no restriction on the contemporaneous correlation between y and m. Moreover, nonzero values of λ_{my} and λ_{ym} allow both y and m to respond ϵ^m and ϵ^η shocks within the period. With the assumption that $\text{cov}(\epsilon_t^m, \epsilon_t^\eta) = 0$, only one additional identifying restriction is required.

Where might this additional restriction come from? One approach is to assume that the model is recursive, so that either $\lambda_{my} = 0$ or $\lambda_{ym} = 0$. Geweke (1986), Stock and Watson (1988), Rotemberg, Driscoll, and Poterba (1995), and Fisher and Seater (1993) present tests for neutrality under the assumption

[5] The unit roots discussion of Section 1 is important here, since the invertability of $\Theta(L)$ requires that $\Theta(1)$ has full rank. This implies that y_t and m_t are both integrated processes, and (y_t, m_t) are not cointegrated.

that $\lambda_{ym} = 0$; Geweke (1986) also presents results under the assumption that $\lambda_{my} = 0$. Alternatively, neutrality might be *assumed*, and the restriction $\gamma_{ym} = 0$ used to identify the model. This assumption has been used by Gali (1992), by King, Plosser, Stock, and Watson (1991), by Shapiro and Watson (1988), and by others to disentangle the structural shocks ϵ_t^m and ϵ_t^η. Finally, an assumption such as $\gamma_{my} = 1$ might be used to identify the model; this assumption is consistent with long-run price stability under the assumption of stable velocity.

The approach that we take in the empirical section is more eclectic and potentially more informative. Rather than report results associated with a single identifying restriction, we summarize results for a wide range of observationally equivalent estimated models. This allows the reader to gauge the robustness of conclusions about γ_{ym} and long-run neutrality to specific assumptions about $\lambda_{ym}, \lambda_{my}$, or γ_{my}. Our method is in the spirit of robustness calculations carried out by sophisticated users of structural VARs such as Sims (1989) and Blanchard (1989).

3. EVIDENCE ON THE NEUTRALITY PROPOSITIONS IN THE POSTWAR U.S. ECONOMY

While our discussion has focused on the long-run neutrality of money, we can test a range of related long-run neutrality propositions by varying the definition X_t in equation (7). As we have shown, using $X_t = (\Delta y_t, \Delta m_t)'$, with m_t assumed to follow an I(1) process, the model can be used to investigate the neutrality of money. If the process describing m_t is I(2) rather than I(1), then the framework can be used to investigate superneutrality by using $X_t = (\Delta y_t, \Delta^2 m_t)'$.[6] In economies in which rate of inflation, π_t, and the nominal interest rate, R_t, follow integrated processes, then we can study the long-run effect of inflation on real interest rates by setting $X_t = (\Delta\pi_t, \Delta R_t)'$. Finally, if both the inflation rate and the unemployment rate are I(1), then the slope of the long-run Phillips curve can be investigated using $X_t = (\Delta\pi_t, \Delta u_t)$.

We investigate these four long-run neutrality hypotheses using postwar quarterly data for the United States. We use gross national product for output;

[6] Long-run neutrality cannot be tested in a system in which output is I(1) and money is I(2). Intuitively this follows because neutrality concerns the relationship between shocks to the level of money and to the level of output. When money is I(2), shocks affect the rate of growth of money, and there are no shocks to the level of money. To see this formally, write equation (8a) as

$$\begin{aligned}\alpha_{yy}(L)\Delta y_t &= \alpha_{ym}(L)\Delta m_t + \epsilon_t^\eta \\ &= \alpha_{ym}(1)\Delta m_t + \alpha_{ym}^*(L)\Delta^2 m_t + \epsilon_t^\eta,\end{aligned}$$

where $\alpha_{ym}^*(L) = (1-L)^{-1}[\alpha_{ym}(L) - \alpha_{ym}(1)]$. When money is I(1), the neutrality restriction is $\alpha_{ym}(1) = 0$. But when money is I(2) and output is I(1), $\alpha_{ym}(1) = 0$ by construction. (When $\alpha_{ym}(1) \neq 0$, output is I(2).) For a more detailed discussion of neutrality restrictions with possibly different orders of integration, see Fisher and Seater (1993).

money is M2; unemployment is the civilian unemployment rate; price inflation is calculated from the consumer price index; and the nominal interest rate is the yield on three-month Treasury bills.[7]

Since the unit root properties of the data play a key role in the analysis, Table 1 presents statistics describing these properties of the data. We use two sets of statistics: (i) augmented Dickey-Fuller (ADF) t-statistics and (ii) 95 percent confidence intervals for the largest autoregressive root. (These were constructed from the ADF statistics using Stock's [1991] procedure.)

The ADF statistics indicate that unit roots cannot be rejected at the 5 percent level for any of the series. From this perspective, output (y_t), money (m_t), money growth (Δm_t), inflation (π_t), unemployment (u_t), and nominal interest rates (R_t) all can be taken to possess the nonstationarity necessary for investigating long-run neutrality using the final form (7). Moreover, a unit root cannot be rejected for $r_t = R_t - \pi_t$, consistent with the hypothesis that R_t and π_t are not cointegrated.

However, the confidence intervals are very wide, suggesting a large amount of uncertainty about the unit root properties of the data. For example, the real GNP data are consistent with the hypothesis that the process is I(1), but also are consistent with the hypothesis that the data are trend stationary with an autoregressive root of 0.89. The money supply data are consistent with the trend stationary, I(1) and I(2) hypotheses. The results in Table 1 suggest that while it is reasonable to carry an empirical investigation of the neutrality propositions predicated on integrated processes, as is usual in models with unit root identifying restrictions, the results must be interpreted with some caution.

Our empirical investigation centers around the four economic interpretations of equation (7) discussed above. For each interpretation, we estimate the model using the following identifying assumptions:

(i) α_0 has 1s on the diagonal,

(ii) Σ_ϵ is diagonal,

and, defining $X_t = (x_t^1\ x_t^2)$, one of the following:

(iii.a) the impact elasticity x^1 with respect to x^2 is known (e.g., λ_{ym} is known in the money-output system),

[7] Data sources: Output: Citibase series GNP82 (real GNP). Money: The monthly Citibase M2 series (FM2) was used for 1959–1989; the earlier M1 data were formed by splicing the M2 series reported in Banking and Monetary Statistics, 1941–1970, Board of Governors of the Federal Reserve System, to the Citibase data in January 1959. Inflation: Log first differences of Citibase series PUNEW (CPI-U: All Items). Unemployment Rate: Citibase Series LHUR (Unemployment rate: all workers, 16 years and over [percent, sa]). Interest Rate: Citibase series FYGM3 (yield on three-month U.S. Treasury bills). Monthly series were averaged to form the quarterly data.

Table 1 Unit Root Statistics

Variable	ADF $\hat{\tau}^{\tau}$	ADF $\hat{\tau}^{\mu}$	95 Percent Confidence Intervals for ρ	
			Detrended Data	Demeaned Data
y_t	−2.53	—	(0.89 1.02)	—
m_t	−2.40	—	(0.90 1.03)	—
Δm_t	−2.76	−2.90	(0.86 1.02)	(0.84 1.01)
π_t	−3.27	−2.86	(0.81 1.02)	(0.84 1.02)
u_t	−3.35	−2.34	(0.81 1.01)	(0.89 1.02)
R_t	−3.08	−1.87	(0.84 1.02)	(0.92 1.02)
r_t	−3.34	−2.94	(0.82 1.02)	(0.85 1.01)

Notes: The regressions used to calculate the ADF statistics included six lagged differences of the variable. All regressions were carried out over the period 1949:1 to 1990:4 using quarterly data except those involving u_t, which began in 1950:1. The variables y_t, m_t are the logarithms of output and money multiplied by 400, so that their first differences represent rates of growth at annual rates; similarly, π_t represents price inflation at an annual rate. The 95 percent confidence intervals were based on the ADF statistics using the procedure developed in Stock (1991).

(iii.b) the impact elasticity of x^2 with respect to x^1 is known (e.g., λ_{my} is known in the money-output system),

(iii.c) the long-run elasticity of x^1 with respect to x^2 is known (e.g., γ_{ym} is known in the money-output system),

(iii.d) the long-run elasticity of x^2 with respect to x^1 is known (e.g., γ_{my} is known in the money-output system).

The models are estimated using simultaneous equation methods. The details are provided in the appendix, but the basic strategy is quite simple and we describe it here using the money-output system. If λ_{ym} in (8a) were known, then the equation could be estimated by regressing $\Delta y_t - \lambda_{ym}\Delta m_t$ onto the lagged values of the variables in the equation. However, the money supply equation (8b) cannot be estimated by ordinary least squares regression since it contains Δy_t, which is potentially correlated with the error term. The maximum likelihood estimator of this equation is constructed by instrumental variables, using the residual from the estimated output supply equation together with lags of Δm_t and Δy_t as instruments. The residual is a valid instrument because of assumption (ii). In the appendix we show how a similar procedure can be used when assumptions (iii.b) – (iii.d) are maintained. Formulae for the standard errors of the estimators are also provided in the appendix.

We report results for a wide range of values of the parameters in assumptions (iii.a) – (iii.d). All of the models include six lags of the relevant variables. The sample period is 1949:1–1990:4 for the models that did not include the unemployment rate; when the unemployment rate was included in the model, the sample period is 1950:1–1990:4. Data prior to the initial periods were used

as lags in the regressions. The robustness of the results to choice of lag length and sample period is discussed below. We now discuss the empirical evidence on the four long-run neutrality propositions.

Neutrality of Money

Figure 1 plots the estimates of the stochastic trends or permanent components in output and money. These were computed as the multivariate Beveridge-Nelson (1981) trends from the estimated bivariate VAR. Also shown in the graph are the NBER business cycle peak and trough dates. Changes in these series at a given date represent changes in the long-run forecasts of output and money associated with the VAR residuals at that date.[8] A scatterplot of these residuals, or innovations in the stochastic trends, is shown in Figure 2. The simple correlation between these innovations is -0.25. Thus, money and output appear to have a negative long-run correlation, at least over this sample period. The important question is the direction of causation explaining this correlation. Simply put, does money cause output or vice versa? This question cannot be answered without an identifying restriction, and we now present results for a range of different identifying assumptions.

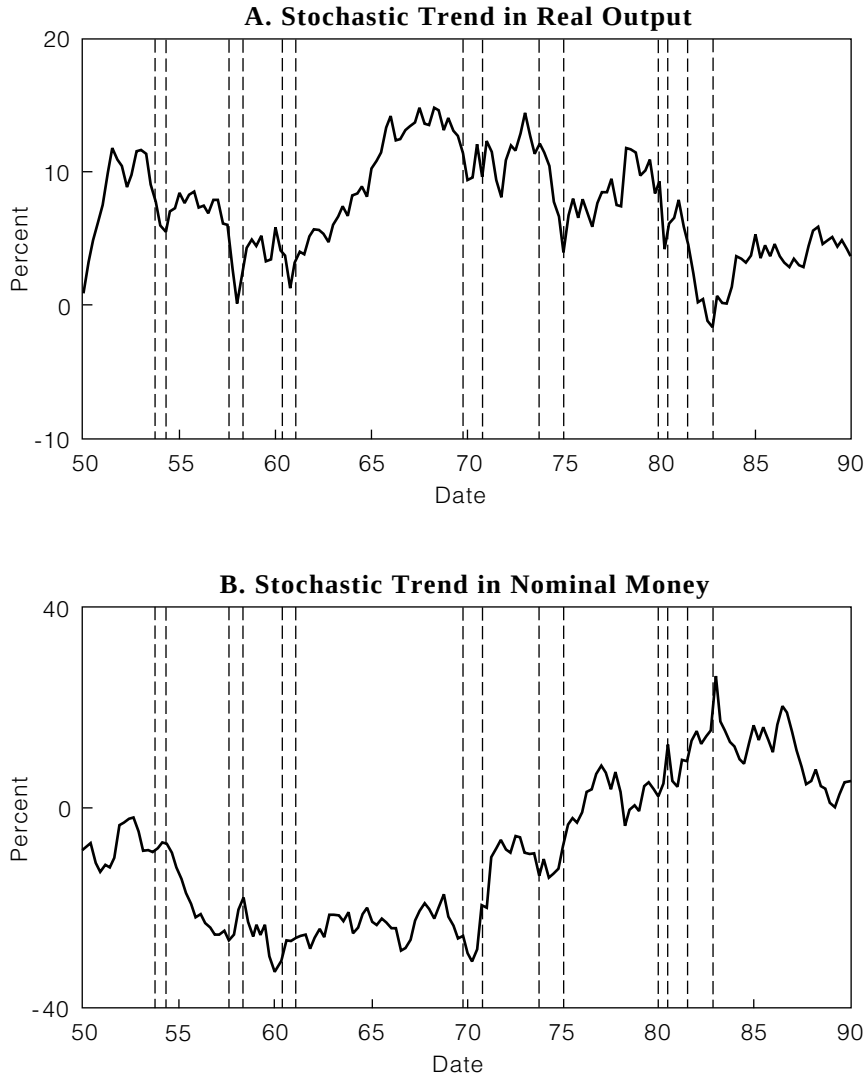
Since we estimate the final form (7) using literally hundreds of different identifying assumptions, there is a tremendous amount of information that can potentially be reported. In Figure 3 we summarize the information on long-run neutrality. Figure 3 presents the point estimates and 95 percent confidence intervals for γ_{ym} for a wide range of values of λ_{my} (panel A), λ_{ym} (panel B), and γ_{my} (panel C). Long-run neutrality is not rejected at the 5 percent level if $\gamma_{ym} = 0$ is contained in the 95 percent confidence interval. For example, from panel A, when $\lambda_{my} = 0$, the point estimate for γ_{ym} is 0.23 and the 95 percent confidence interval is $-0.18 \leq \gamma_{ym} \leq 0.64$. Thus, when $\lambda_{my} = 0$, the data do not reject the long-run neutrality hypothesis. Indeed, as is evident from the figure, long-run neutrality cannot be rejected at the 5 percent level for any value of $\lambda_{my} \leq 1.40$. Thus, the interpretation of the evidence on long-run neutrality depends critically on the assumed value of λ_{my}.

The precise value of λ_{my} depends on the money supply process. For example, if the central bank's reserve position is adjusted to smooth interest rates, then m_t will adjust to accommodate shifts in money demand arising from changes in y_t. In this case, λ_{my} corresponds to the short-run elasticity of money demand, and a reasonable range of values is $0.1 \leq \lambda_{my} \leq 0.6$. For all values of λ_{my} in this range, the null hypothesis of long-run neutrality cannot be rejected.

Panel B of Figure 3 shows that long-run neutrality is not rejected for values of $\lambda_{ym} > -4.61$. Since traditional monetary models of the business cycle imply

[8] Because the VAR residuals sum to zero over the entire sample, the trends are constrained to equal zero in the final period. In addition, they are normalized to equal zero in the initial period. This explains their "Brownian Bridge" behavior.

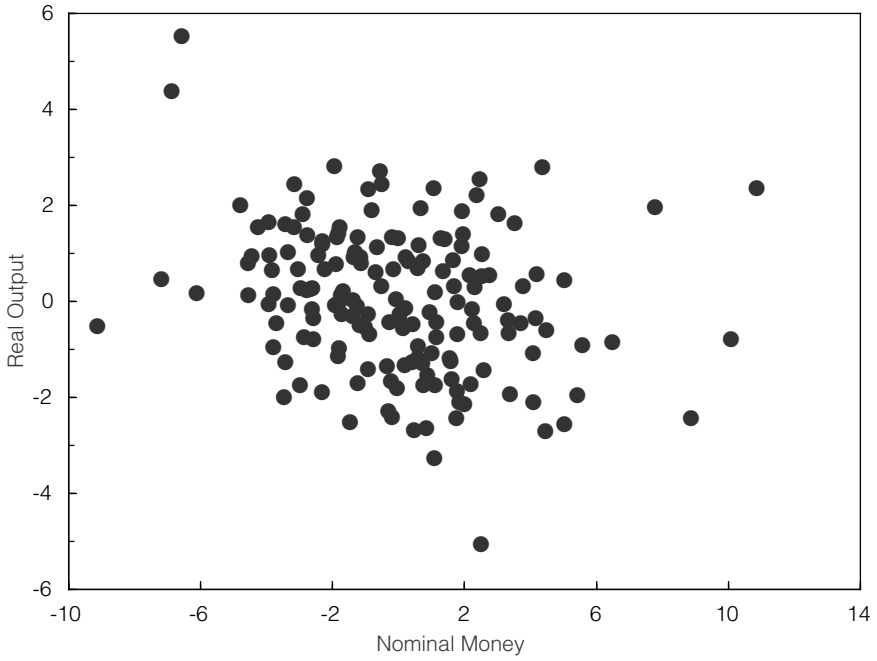
Figure 1 Stochastic Trends



that $\lambda_{ym} \geq 0$—output does not decline on impact in response to a monetary expansion—the results in panel B again suggest that the data are consistent with the long-run neutrality hypothesis.

Finally, the results in panel C suggest that the long-run neutrality hypothesis cannot be rejected for the entire range of values γ_{my} shown in Figure 3. To interpret the results in this figure, recall that γ_{my} represents the long-run response

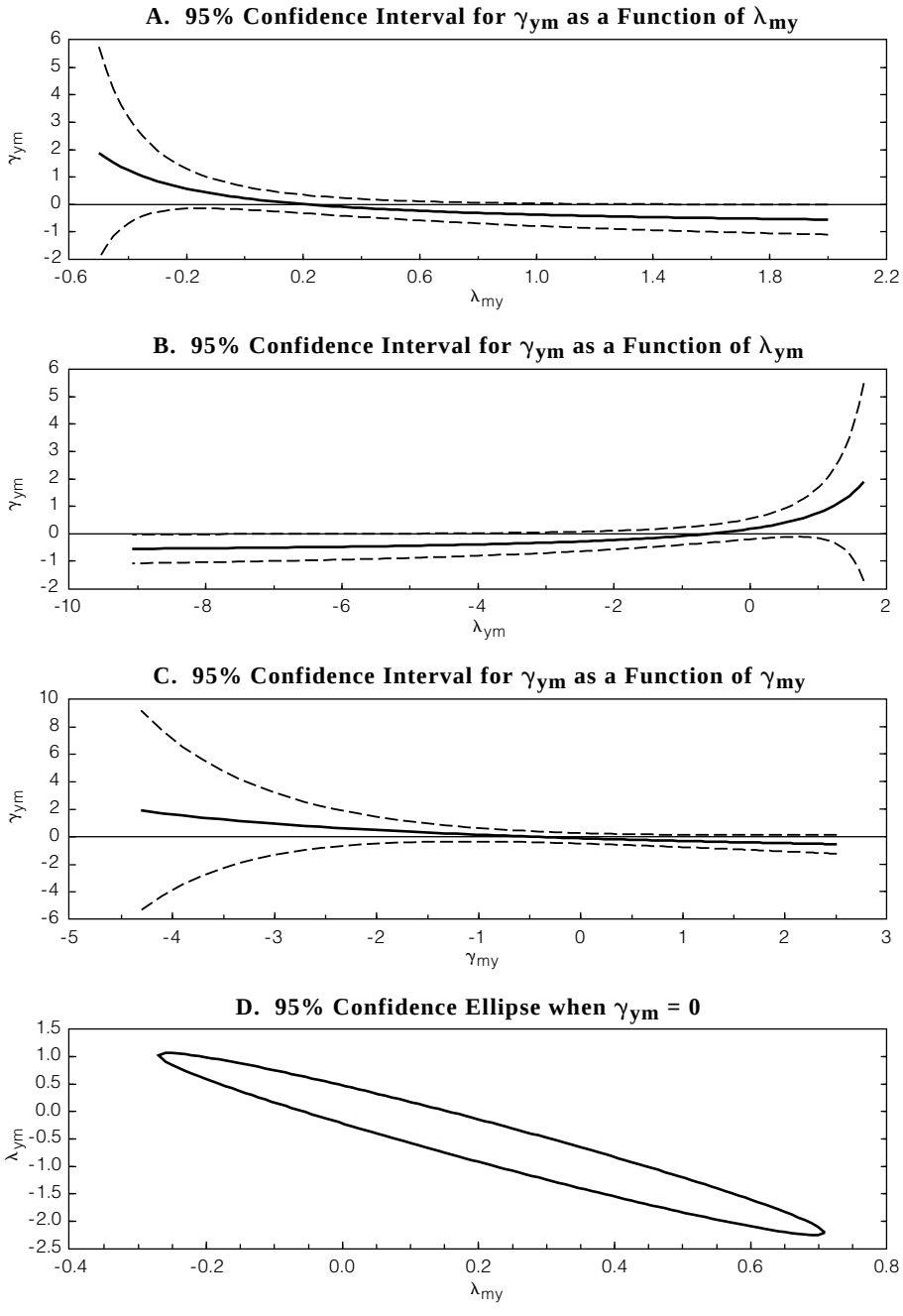
Figure 2 Innovations in Stochastic Trends



of m_t to exogenous permanent shifts in the level of y_t. If (M2) velocity is reasonably stable over long periods, then price stability would require $\gamma_{my} = 1$. Consequently, values of $\gamma_{my} < 1$ represent long-run deflationary policies and $\gamma_{my} > 1$ represent long-run inflationary policies. Thus, when $\gamma_{my} = 1 + \delta$, the long-run level of prices increase by δ percent when the long-run level of output increases by 1 percent. In the figure we show that long-run neutrality cannot be rejected for values of γ_{my} as large as 2.5; we have estimated the model using values of γ_{my} as large as 5.7 and found no rejections of the long-run neutrality hypothesis.

An alternative way to interpret the evidence from panels A–C of Figure 3 is to use long-run neutrality as an identifying restriction and to estimate the other parameters of the model. From the figure, when $\gamma_{ym} = 0$, the point estimates are $\hat{\lambda}_{my} = 0.22$, $\hat{\lambda}_{ym} = -0.59$, and $\hat{\gamma}_{my} = -0.51$, and the implied 95 percent confidence intervals are $-0.18 \leq \lambda_{my} \leq 0.62$, $-1.93 \leq \lambda_{ym} \leq 0.74$, and $-2.1 \leq \gamma_{my} \leq 1.06$. By definition, these intervals contain the true values of $\lambda_{my}, \lambda_{ym}$, and γ_{my} 95 percent of the time, if long-run neutrality is true. Thus, if the confidence intervals contain only nonsensical values of these parameters, then this provides evidence against long-run neutrality. We find that the

Figure 3 Money and Output



confidence intervals include many reasonable values of the parameters and conclude that they provide little evidence against the neutrality hypothesis.

Multivariate confidence intervals can also be constructed. Panel D of Figure 3 provides an example. It shows the 95 percent confidence ellipse for (λ_{my}, λ_{ym}) constructed under the assumption of long-run neutrality.[9] If long-run neutrality holds, then 95 percent of the time this ellipse will cover the true values of the pair (λ_{ym}, λ_{my}). Thus, if reasonable values for the pair of parameters are not included in this ellipse, then this provides evidence against long-run neutrality.

Table 2 summarizes selected results for variations in the specification. The VAR lag length (6 in the results discussed above) is varied between 4 and 8, and the model is estimated over various subsamples. Overall, the table suggests that the results are robust to these changes in the specification.[10]

These conclusions are predicated on the two-shock model that forms the basis of the bivariate specification. That is, the analysis is based on the assumption that money and output are driven by only two structural disturbances, here interpreted as a monetary shock and a real shock. This is clearly wrong, as there are many sources of real shocks (productivity, oil prices, tax rates, etc.) and nominal shocks (factors affecting both money supply and money demand). However, deducing the effects of these omitted variables on the analysis is difficult, since what matters is both the relative variability of these different shocks and their different dynamic effects on y and m. Indeed, as shown in Blanchard and Quah (1989), a two-shock model will provide approximately correct answers if the dynamic responses of y and m to shocks with large relative variances are sufficiently similar.

Superneutrality of Money

Evidence on the superneutrality of money is summarized in Figure 4 and in panel B of Table 2. Figure 4 is read the same way as Figure 3, except that now the experiment involves the effects of changes in the rate of growth of

[9] This confidence ellipse is computed in the usual way. For example, see Johnston (1984), p. 190.

[10] These results are not robust to certain other changes in the specification. For example, Rotemberg, Driscoll, and Poterba (1995) report results using monthly data on M2 and U.S. Industrial Production (IP) for a specification that includes a linear time trend, 12 monthly lags, and is econometrically identified using the restriction that $\lambda_{my} = 0$. These authors report an estimate of $\gamma_{ym} = 1.57$ that is significantly different from zero and thus reject long-run neutrality. Stock and Watson (1988) report a similar finding using monthly data on IP and M1. The sample period and output measure seems to be responsible for the differences between these results and those reported here. For example, assuming $\lambda_{ym} = 0$ and using quarterly IP and M2 results in estimated values of γ_{ym} of 0.43 (0.31) using data from 1949:1 to 1990:4. (The standard error of the estimate is shown in parentheses.) As in Table 2, when the sample is split and the model estimated over the period 1949:1 to 1972:4 and 1973:1 to 1990:4, the resulting estimates are 0.56 (0.37) and 1.32 (0.70). Thus, point estimates of γ_{ym} are larger using IP in place of real GNP, and tend to increase in the second half of the second period.

Table 2 Robustness to Sample Period and Lag Length

A. Neutrality of Money
$X_t = (\Delta m_t,\ \Delta y_t)'$

Sample Period	Lag Length	Estimates of γ_{ym} when $\lambda_{my} = 0$		Estimates of γ_{ym} when $\lambda_{ym} = 0$		Estimates of γ_{ym} when $\gamma_{my} = 0$	
1949–1990	6	0.23	(0.21)	0.17	(0.19)	−0.12	(0.19)
1949–1972	6	0.15	(0.24)	0.13	(0.24)	0.04	(0.27)
1973–1990	6	0.77	(0.47)	0.65	(0.37)	0.02	(0.25)
1949–1990	4	0.24	(0.17)	0.20	(0.15)	−0.04	(0.17)
1949–1990	8	0.12	(0.19)	0.07	(0.17)	−0.18	(0.18)

B. Superneutrality of Money
$X_t = (\Delta^2 m_t,\ \Delta y_t)'$

Sample Period	Lag Length	Estimates of γ_y, Δ_m when $\lambda_{\Delta m,y} = 0$		Estimates of γ_y, Δ_m when $\lambda_{y,\Delta m} = 0$		Estimates of γ_y, Δ_m when $\gamma_{\Delta m,y} = 0$	
1949–1990	6	3.80	(1.74)	3.12	(1.36)	−0.95	(1.57)
1949–1972	6	3.50	(1.66)	3.32	(1.49)	1.67	(1.99)
1973–1990	6	4.02	(4.57)	2.65	(2.62)	−4.11	(1.14)
1949–1990	4	1.81	(0.90)	1.31	(0.63)	−1.55	(0.97)
1949–1990	8	3.94	(1.81)	3.43	(1.53)	0.10	(1.66)

C. Long-Run Fisher Effect
$X_t = (\Delta \pi_t,\ \Delta R_t)'$

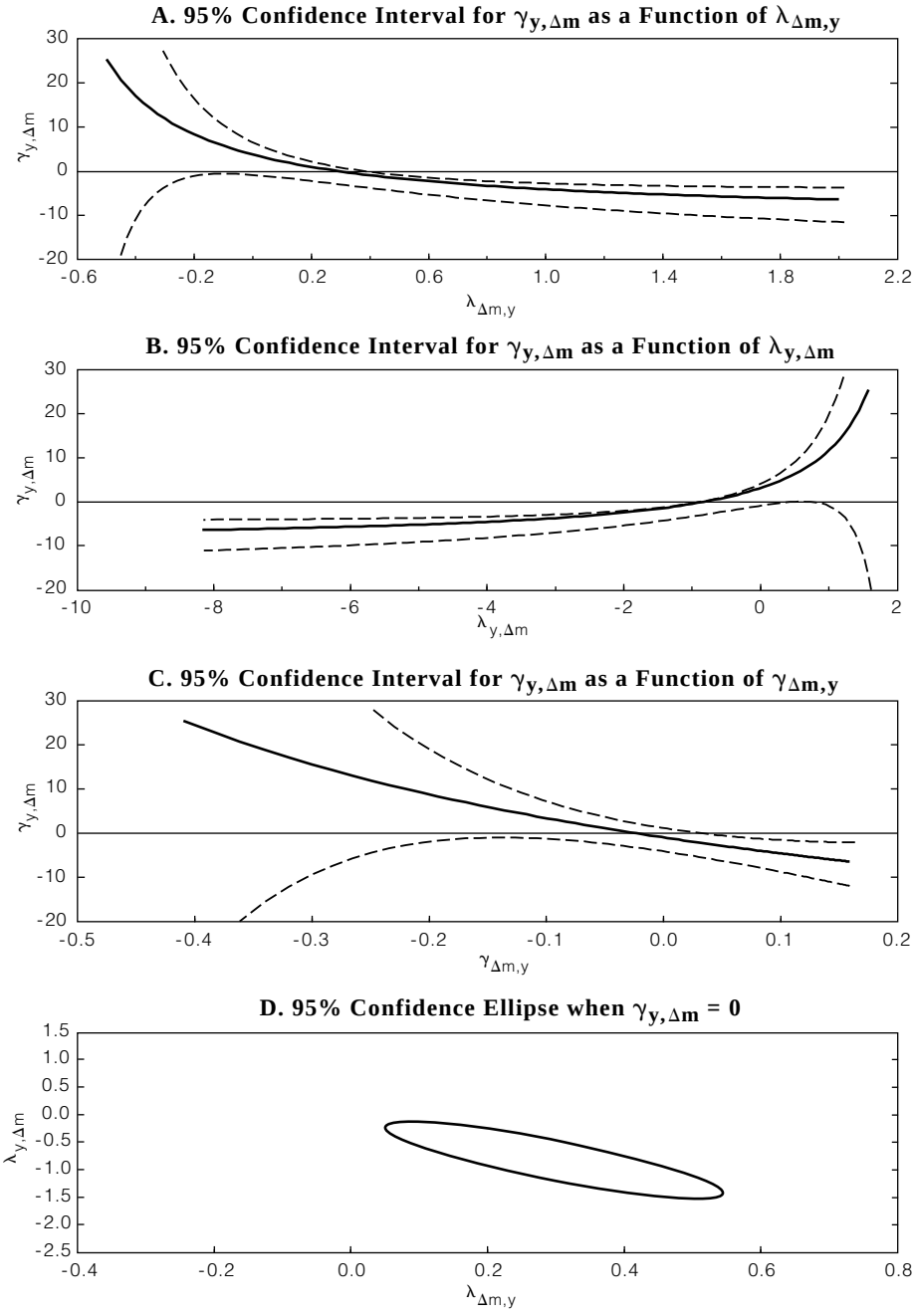
Sample Period	Lag Length	Estimates of $\gamma_{R\pi}$ when $\lambda_{\pi R} = 0$		Estimates of $\gamma_{R\pi}$ when $\lambda_{R\pi} = 0$		Estimates of $\gamma_{R\pi}$ when $\gamma_{\pi R} = 0$	
1949–1990	6	0.18	(0.09)	0.08	(0.08)	0.34	(0.12)
1949–1972	6	0.04	(0.06)	0.03	(0.05)	0.07	(0.09)
1973–1990	6	0.40	(0.16)	0.23	(0.18)	0.53	(0.20)
1949–1990	4	0.15	(0.07)	0.07	(0.06)	0.28	(0.09)
1949–1990	8	0.26	(0.09)	0.14	(0.08)	0.39	(0.13)

D. Long-Run Phillips Curve
$X_t = (\Delta \pi_t,\ \Delta u_t)'$

Sample Period	Lag Length	Estimates of $\gamma_{u\pi}$ when $\lambda_{\pi u} = 0$		Estimates of $\gamma_{u\pi}$ when $\lambda_{u\pi} = 0$		Estimates of $\gamma_{u\pi}$ when $\gamma_{\pi u} = 0$	
1950–1990	6	0.03	(0.09)	0.06	(0.09)	−0.17	(0.11)
1950–1972	6	−0.04	(0.10)	−0.03	(0.09)	−0.07	(0.14)
1973–1990	6	0.29	(0.35)	0.51	(0.56)	−0.21	(0.16)
1950–1990	4	−0.03	(0.06)	−0.00	(0.05)	−0.18	(0.07)
1950–1990	8	0.08	(0.09)	0.12	(0.09)	−0.11	(0.10)

Note: Standard errors are shown in parentheses.

Figure 4 Money Growth and Output



money, so that the parameters are $\lambda_{\Delta m,y}$, $\lambda_{y,\Delta m}$, $\gamma_{\Delta m,y}$, and $\gamma_{y,\Delta m}$. There are two substantive conclusions to be drawn from the table and figure.

The first conclusion is that it is possible to find evidence against superneutrality. For example, superneutrality is rejected at the 5 percent level for all values of $\lambda_{\Delta m,y}$ between -0.25 and 0.08, and for all values of $\lambda_{y,\Delta m}$ between -0.26 and 1.02. On the other hand, the figures suggest that these rejections are marginal, and the rejections are not robust to all of the lag-length and sample-period specification changes reported in Table 2. Moreover, a wide range of (arguably) reasonable identifying restrictions lead to the conclusion that superneutrality cannot be rejected. For example, superneutrality is not rejected for any value of $\lambda_{\Delta m,y}$ in the interval 0.08 to 0.53. Because of the lags in the model, the impact multiplier $\lambda_{\Delta m,y}$ has the same interpretation as λ_{my} in the discussion of long-run neutrality, and we argued above that the interval (0.08, 0.53) was a reasonable range of values for this parameter. In addition, from panel C, superneutality cannot be rejected for values of $\gamma_{\Delta m,y} < 0.07$. To put this into perspective, note that $\gamma_{\Delta m,y}$ measures the long-run elasticity of rate of growth of money with respect to permanent changes in the level of output. Thus a value of $\gamma_{\Delta m,y} = 0$ corresponds to a non-accelerationist policy.

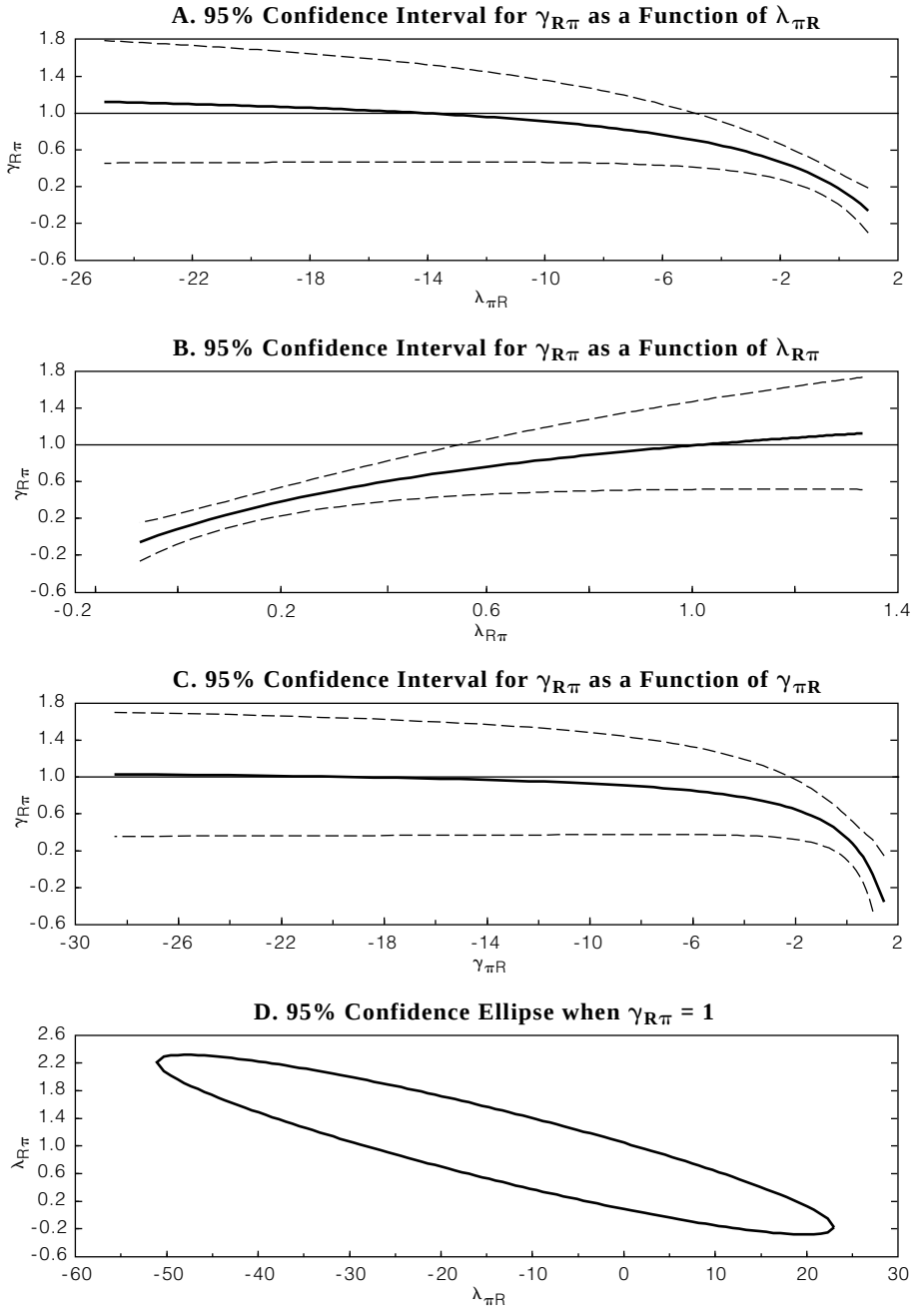
The second substantive conclusion is that the identifying assumption has a large effect on the sign and the magnitude of the estimated value of $\gamma_{y,\Delta m}$. For example, when $\lambda_{\Delta m,y} = 0$ the estimated value of $\gamma_{y,\Delta m}$ is 3.8. Thus, a 1 percent permanent increase in the money growth rate is estimated to increase the flow of output by 3.8 percent per year in perpetuity. Our sense is that even those who believe that the Tobin (1965) effect is empirically important do not believe that it is this large. The estimated value of $\gamma_{y,\Delta m}$ falls sharply as $\lambda_{\Delta m,y}$ is increased, and $\hat{\gamma}_{y,\Delta m} = 0$ when $\lambda_{\Delta m,y} = 0.30$. For values of $\lambda_{\Delta m,y} > 0.30$, the point estimate of $\gamma_{y,\Delta m}$ is negative, consistent with the predictions of cash-in-advance models in which sustained inflation is a tax on investment activity (Stockman 1981) or on labor supply (Aschauer and Greenwood 1983 or Cooley and Hansen 1989).

The Fisherian Theory of Inflation and Interest Rates

In the Fisherian theory of interest, the interest rate is determined as the sum of a real component, r_t, and an expected inflation component $E_t\pi_{t+1}$. A related long-run neutrality proposition—also suggested by Fisher—is that the level of the real interest rate is invariant to permanent changes in the rate of inflation. If inflation is integrated, then this proposition can be investigated using our framework: when $X_t = (\Delta\pi_t, \Delta R_t)$, then permament changes in π_t will have no effect on real interest rates when $\gamma_{R\pi} = 1$.

We find mixed evidence against the classical Fisherian link between long-run components of inflation and nominal interest rates, interpreted here as $\gamma_{R\pi} = 1$. For example, from Figure 5, maintaining a positive value of either

Figure 5 Inflation and Nominal Rates



$\lambda_{\pi R}$ or $\gamma_{\pi R}$ leads to an estimate of $\gamma_{R\pi}$ that is significantly less than 1. A mechanical explanation of this finding is that the VAR model implies substantial volatility in trend inflation: the estimated standard deviation of the inflation trend is much larger (1.25) than that of nominal rates (0.75). Thus, to reconcile the data with $\gamma_{R\pi} = 1$, a large negative effect of nominal interest rates on inflation is required.

However, from panel B of the figure, $\gamma_{R\pi} = 1$ cannot be rejected for a value of $\lambda_{R\pi} > 0.55$. One way to interpret the $\lambda_{R\pi}$ parameter is to decompose the impact effect of π on R into an expected inflation effect and an effect on real rates. If π has no impact effect on real rates, so that only the expected inflation effect was present, then $\lambda_{R\pi} = \partial\pi_{t+1}/\partial\epsilon_t^{\pi}$. For our data, $\partial\pi_{t+1}/\partial\epsilon_t^{\pi} = 0.6$ when the model is estimated using $\lambda_{R\pi} = 0.6$ as an identifying restriction, suggesting that this is a reasonable estimate of the expected inflation effect. The magnitude of the real interest effect is more difficult to determine since different macreconomic models lead to different conclusions about the effect of nominal shocks on real rates. For example, models with liquidity effects imply that real rates fall (e.g., Lucas [1990], Fuerst [1992], and Christiano and Eichenbaum [1994]), while the sticky nominal wage and price models in King (1994) imply that real rates rise. In this regard, the interpretation of the evidence on the long-run Fisher effect is seen to depend critically on one's belief about the impact effect of a nominal disturbance on the real interest rate. If this effect is negative, then there is significant evidence in the data against this neutrality hypothesis.

The confidence intervals suggest that the evidence against the long-run Fisher relation is not overwhelming. When $\gamma_{R\pi} = 1$ is maintained, the implied confidence intervals for the other parameters are wide ($-43.7 \leq \lambda_{\pi R} \leq 15.6, 0.0 \leq \lambda_{R\pi} \leq 2.1, -154.8 \leq \gamma_{\pi R} \leq 116.4$) and contain what are arguably reasonable values of these parameters. This is also evident from the confidence ellipse in panel D of Figure 5.

One interpretation is that these results reflect the conventional finding that nominal interest rates do not adjust fully to sustained inflation in the postwar U.S. data. This result obtains for a wide range of identifying assumptions. One possible explanation is that the failure depends on the particular specification of the bivariate model that we employ, suggesting the importance of extending this analysis to multivariate models. Another candidate source of potential misspecification is cointegration between nominal rates and inflation. This is discussed in some detail in papers by Evans and Lewis (1993), Mehra (1995), and Mishkin (1992).[11]

[11] These authors suggest that real rates $R_t - \pi_t$ are I(0). Evans and Lewis (1993) and Mishkin (1992) find estimates suggesting that nominal rates do not respond fully to permanent changes in inflation and attribute this to a small sample bias associated with shifts in the inflation process. Mehra (1995) finds that permanent changes in interest rates do respond one-for-one with

4. EVIDENCE ON THE LONG-RUN PHILLIPS CURVE

As discussed in King and Watson (1994), the interpretation of the evidence on the long-run Phillips curve is more subtle than the other neutrality propositions.[12] Throughout this article we have examined neutrality by examining the long-run multiplier in equations relating real variables to nominal variables. This suggests examining the neutrality proposition embodied in the long-run Phillips curve using the equation

$$\alpha_{uu}(L)u_t = \alpha_{u\pi}(L)\pi_t + \epsilon_t^u. \tag{12}$$

Of course, as in Sargent (1976), equation (12) is one standard way of writing the Phillips curve.

Figure 6 shows estimates $\gamma_{u\pi}$ for a wide range of identifying assumptions. When the model is estimated using $\lambda_{\pi u}$ as an identifying assumption, a vertical Phillips curve ($\gamma_{u\pi} = 0$) is rejected when $\lambda_{\pi u} > 2.3$.[13] Thus, neutrality is rejected only if one assumes that positive changes in the unemployment rate have a large positive impact effect on inflation. From panel B of the figure, $\gamma_{u\pi} = 0$ is rejected for maintained values of $\lambda_{u\pi} < -0.07$. Since $\lambda_{u\pi}$ can be interpreted as the slope of the short-run (impact) Phillips curve, this figure shows the relationship between maintained assumptions and conclusions about short-run and long-run neutrality. The data are consistent with the pair of parameters $\lambda_{u\pi}$ and $\gamma_{u\pi}$ being close to zero; the data also are consistent with the hypothesis that these parameters are both less than zero. If short-run neutrality is maintained ($\lambda_{u\pi} = 0$), the estimated long-run effect of inflation on unemployment is very small ($\hat{\gamma}_{u\pi} = 0.06$). If long-run neutrality is maintained ($\gamma_{u\pi} = 0$), the estimated short-run effect of inflation on unemployment is very small ($\hat{\lambda}_{u\pi} = -0.02$). This latter result is consistent with the small estimated real effects of nominal disturbances found by King, Plosser, Stock, and Watson (1991), Gali (1992), and Shapiro and Watson (1988), who all used long-run neutrality as an identifying restriction.

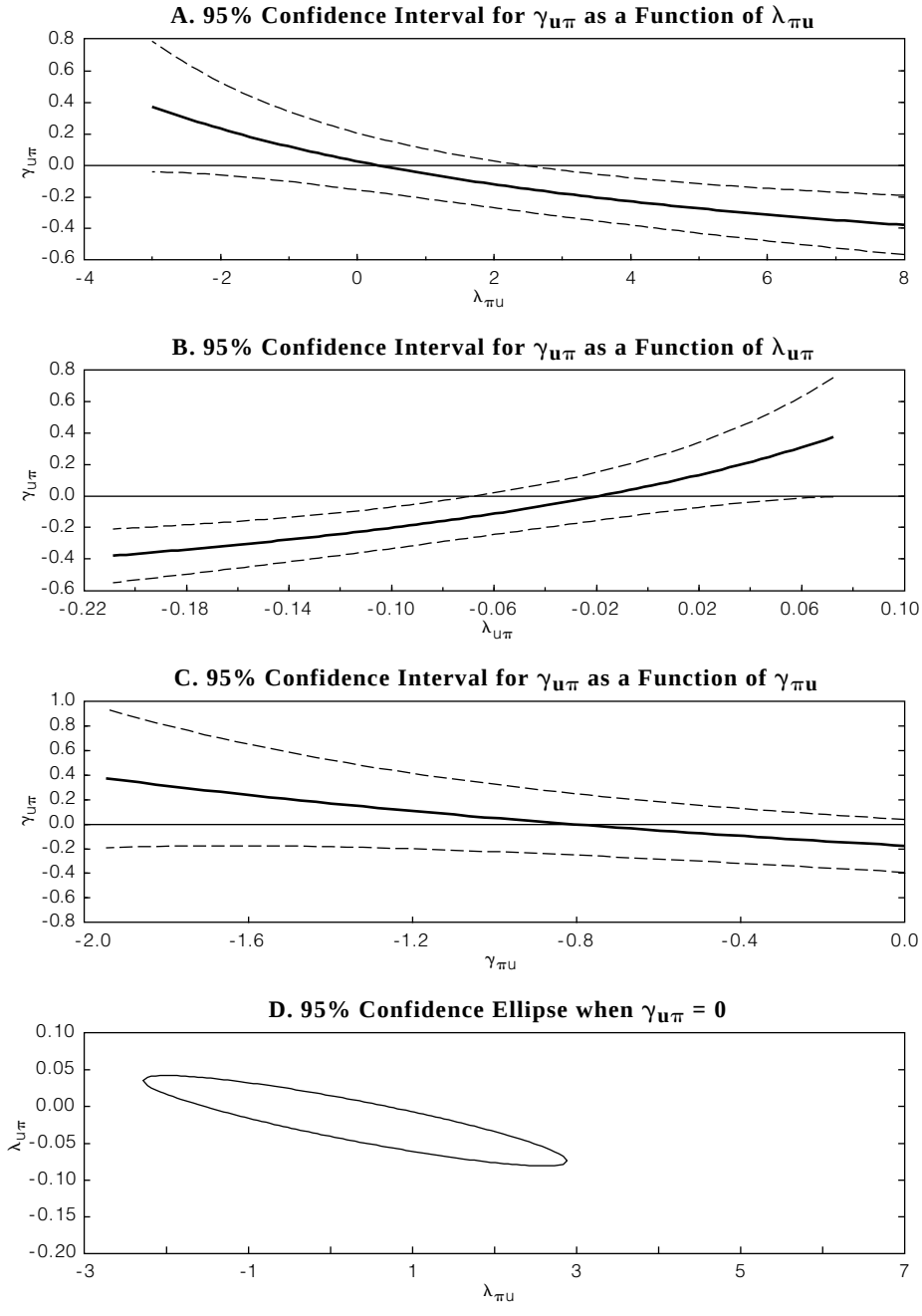
Several researchers, relying on a variety of specifications and identifying assumptions, have produced estimates of the short-run Phillips curve slope. For example, Sargent (1976) estimates $\lambda_{u\pi}$ using innovations in population, money, and various fiscal policy variables as instruments. He finds an estimate of $\lambda_{u\pi} = -0.07$. Estimates of $\lambda_{u\pi}$ ranging from -0.07 to -0.18 can

permanent changes in inflation. In contrast to these papers, our results are predicated on the assumption that π_t and R_t are I(1) and are not cointegrated over the entire sample. As the results in Table 1 make clear, both the I(0) and I(1) hypotheses are consistent with the data.

[12] A greatly expanded version of the analysis in this section is contained in King and Watson (1994).

[13] Recall that the Phillips curve is drawn with inflation on the vertical axis and unemployment on the horizontal axis. Thus, a vertical long-run Phillips curve corresponds to the restriction $\gamma_{u\pi} = 0$.

Figure 6 Inflation and Unemployment



be extracted from the results in Barro and Rush (1980), who estimated the unemployment and inflation effects of unanticipated money shocks. Values of $\lambda_{u\pi}$ in this range lead to a rejection of the null $\gamma_{u\pi} = 0$, but they suggest a very steep long-run tradeoff. For example, when $\lambda_{u\pi} = -0.10$, the corresponding point estimate of $\gamma_{u\pi} = -0.20$, so that the long-run Phillips curve has a slope of $-5.0(= \gamma_{u\pi}^{-1})$.

By contrast, the conventional view in the late 1960s and early 1970s was that there was a much more favorable tradeoff between inflation and unemployment. For example, in discussing Gordon's famous (1970) test of an accelerationist Phillips curve model, Solow calculated that there was a one-for-one long-run tradeoff implied by Gordon's results. This calculation was sufficiently conventional that it led to no sharp discussion among the participants at the Brookings panel. Essentially the same tradeoff was suggested by the 1969 *Economic Report of the President,* which provided a graph of inflation and unemployment between 1954 and 1968.[14]

What is responsible for the difference between our estimates and the conventional estimates from the late '60s? Panel D in Table 2 suggests that sample period cannot be the answer: the full sample results are very similar to the results obtained using data from 1950 through 1972. Instead, the answer lies in differences between the identifying assumptions employed. The traditional Gordon-Solow estimate was obtained from a price equation of the form[15]

$$\alpha_{\pi\pi}(L)\pi_t = \alpha_{\pi u}(L)u_t + \epsilon_t^{\pi}. \tag{13}$$

The estimated slope of the long-run Phillips curve was calculated as $\gamma = \alpha_{\pi u}(1)/\alpha_{\pi\pi}(1)$. Thus, in the traditional Gordon-Solow framework, the long-run Phillips curve was calculated as the long-run multiplier from the inflation equation. In contrast, our estimate ($\gamma_{u\pi}^{-1}$) is calculated from the unemployment equation. The difference is critical, since it means that the two parameters represent responses to different shocks. Using our notation, the long-run multiplier from (13) is

$$\gamma_{\pi u} = \frac{\lim_{k\to\infty} \partial\pi_{t+k}/\partial\epsilon_t^u}{\lim_{k\to\infty} \partial u_{t+k}/\partial\epsilon_t^u},$$

while the inverse of the long-run multiplier from the unemployment equation (12) is

$$\gamma_{u\pi}^{-1} = \frac{\lim_{k\to\infty} \partial\pi_{t+k}/\partial\epsilon_t^{\pi}}{\lim_{k\to\infty} \partial u_{t+k}/\partial\epsilon_t^{\pi}}.$$

[14] See McCallum (1989, p. 180) for a replication and discussion of this graph.

[15] Equation (13) served as a baseline model for estimating the Phillips curve. Careful researchers employed various shift variables in the regression to capture the effects of demographic shifts on the unemployment rate and the effects of price controls on inflation. For our purposes, these complications can be ignored.

Thus, the traditional estimate measures the relative effect of shocks to unemployment, while our estimate corresponds to the relative effect of shocks to inflation. Figure 7 presents our estimates of $\gamma_{\pi u}$. Evidently, the Gordon-Solow value of $\gamma_{u\pi} = -1$ is consistent with a wide range of identifying restrictions shown in the figure.

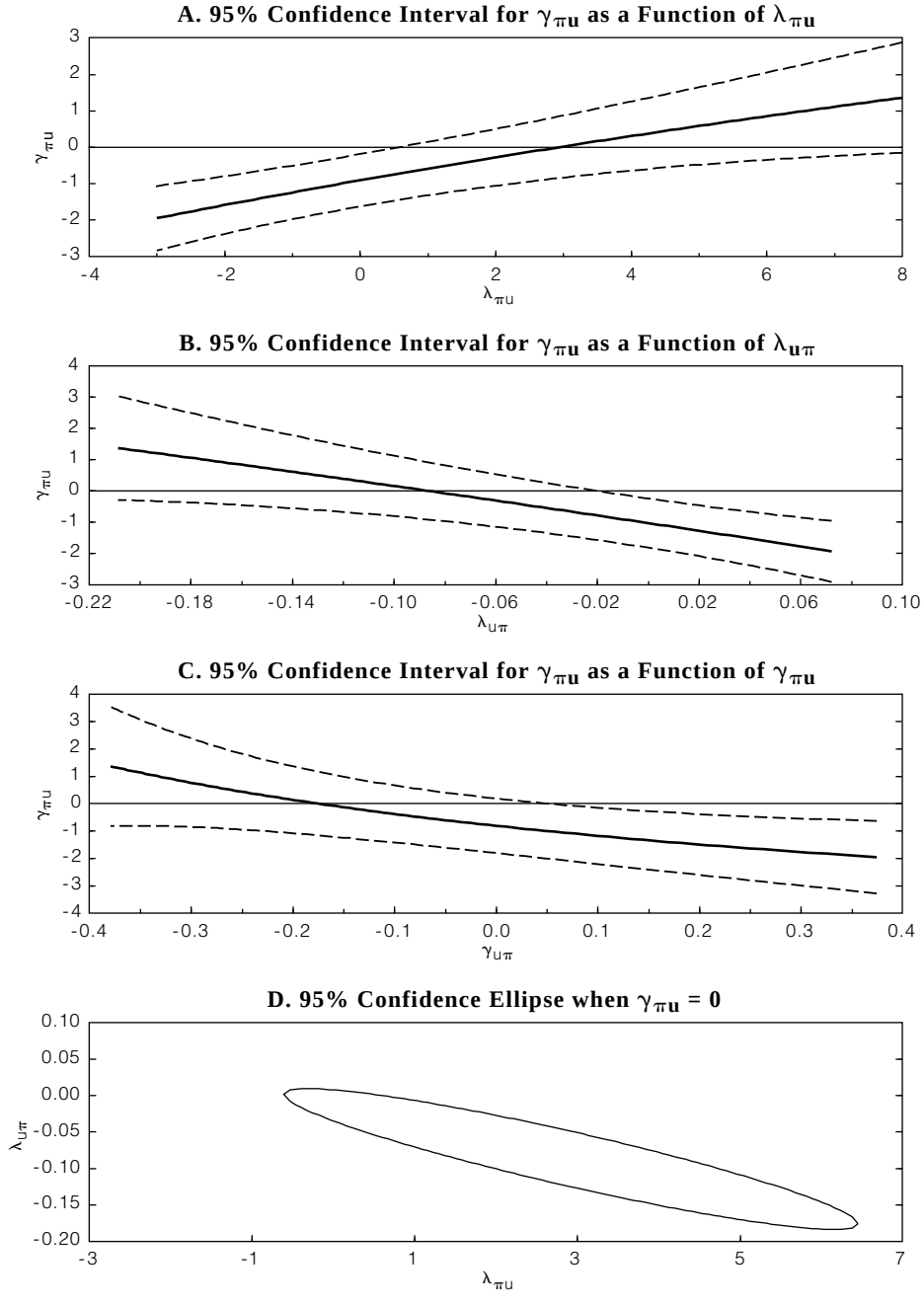
But the question is not whether the long-run multiplier is calculated from the unemployment equation, $\alpha_{uu}(L)u_t = \alpha_{u\pi}(L)\pi_t + \epsilon_t^u$, or from the inflation equation, $\alpha_{\pi\pi}(L)\pi_t = \alpha_{\pi u}(L)u_t + \epsilon_t^\pi$. By choosing between these two specifications under a specific identification scheme, one is also choosing a way of representing the experiment of a higher long-run rate of inflation, presumably originating from a higher long-run rate of monetary expansion. Under the Gordon-Solow procedure, the idea is that the shock to unemployment—the ϵ_t^u shock defined by a particular identifying restriction—is the indicator of a shift in aggregate demand. Its consequences are traced through the inflation equation since unemployment is the right-hand side variable in that equation. Under the Lucas-Sargent procedure, the idea is that the shock to inflation—the ϵ_t^π shock defined by a particular identifying restriction—is the indicator of a shift in aggregate demand.

To interpret the Gordon-Solow estimate of $\gamma_{\pi u}$ we must determine the particular identifying assumption that they used. Their assumption can be deduced from the way that they estimated $\gamma_{\pi u}$, namely from the ordinary least squares estimators of equation (13). Recall that OLS requires that the variables on the right-hand side of (13) are uncorrelated with the error term. Since u_t appears on the right-hand side of (13), this will be true only when $\lambda_{u\pi} = 0$. Thus, the particular identifying assumption employed in the Gordon-Solow specification in $\lambda_{u\pi} = 0$.

What does this identifying assumption mean? When $\lambda_{u\pi} = 0$, the Gordon-Solow interpretation implies that autonomous shocks to aggregate demand are one-step-ahead forecast errors in u_t. The other shocks in the system can affect prices on impact but cannot affect unemployment. Thus, in this sense, prices are flexible, since they can be affected on impact by all shocks, but unemployment is sticky, since it can be affected on impact only by aggregate demand shocks. For today's "new Keynesians" this may appear to be a very unreasonable identifying restriction (and so must any evidence about the Phillips curve that follows from it). However, the identifying restriction is consistent with the traditional Keynesian model of the late 1960s.[16]

[16] What we have in mind is a block recursive model in which the unemployment rate is determined in an IS-LM block, and wages and prices are determined in a wage-price block. This interpretation is further explored in King and Watson (1994).

Figure 7 Unemployment and Inflation



5. CONCLUDING REMARKS

We have investigated four long-run neutrality propositions using bivariate models and 40 years of quarterly observations. We conclude that the data contain little evidence against the long-run neutrality of money and suggest a very steep long-run Phillips curve. These conclusions are robust to a wide range of identifying assumptions. Conclusions about the long-run Fisher effect and the superneutrality of money are not robust to the particular identifying assumption. Over a fairly broad range of identifying restrictions, the data suggest that nominal interest rates do not move one-for-one with permanent shifts in inflation. The sign and magnitude of the estimated long-run effect of money growth on the level of output depends critically on the specific identifying restriction employed.

These conclusions are tempered by four important caveats. First, the results are predicated on specific assumptions concerning the degree of integration of the data, and with 40 years of data the degree of integration is necessarily uncertain. Second, even if the degree of integration were known, only limited "long-run" information is contained in data that span 40 years. This suggests that a useful extension of this work is to carry out similar analyses on long annual series. Third, the analysis has been carried out using bivariate models. If there are more than two important sources of macroeconomic shocks, then bivariate models may be subject to significant omitted variable bias. Thus another extension of this work is to expand the set of variables under study to allow a richer set of structural macroeconomic shocks. The challenge is to do this in a way that produces results that can be easily interpreted in spite of the large number of identifying restrictions required. Fourth, we have analyzed each of these propositions separately and yet there are obvious and important theoretical connections between them. Future work on multivariate extensions of this approach may allow for a unified econometric analysis of these long-run neutrality propositions.

APPENDIX

Estimation Methods

Under each alternative identifying restriction, the Gaussian maximum likelihood estimates can be constructed using standard regression and instrumental variable calculations. When λ_{ym} is assumed known, equation (8a) can be estimated by ordinary least squares by regressing $\Delta y_t - \lambda_{ym}\Delta m_t$ onto $\{\Delta y_{t-i}, \Delta m_{t-i}\}_{i=1}^{p}$. Equation (8b) cannot be estimated by OLS because Δy_t, one of the regressors, is potentially correlated with ϵ_t^m. Instrumental variables must be used. The appropriate instruments are $\{\Delta y_{t-i}, \Delta m_{t-i}\}_{i=1}^{p}$ together with the residual from the estimated (8a). This residual is a valid instrument because of the assumption that ϵ_t^{η} and ϵ_t^m are uncorrelated. When λ_{my} is assumed known, rather than λ_{ym}, this process was reversed.

When a value for γ_{my} is used to identify the model, a similar procedure can be used. First, rewrite (8b) as

$$\Delta m_t = \alpha_{my}(1)\Delta y_t + \beta_{mm}\Delta m_{t-1} + \sum_{j=0}^{p-1} \tilde{\alpha}_{my}^{j}\Delta^2 y_{t-j} + \sum_{j=0}^{p-1} \tilde{\alpha}_{mm}^{j}\Delta^2 m_{t-j} + \epsilon_t^m, \tag{A1}$$

where $\beta_{mm} = \sum_{j=1}^{p} \alpha_{mm}^{j}$. Equation (A1) replaces the regressors $(\Delta y_t, \Delta y_{t-1}, \ldots, \Delta y_{t-p}, \Delta m_{t-1}, \ldots, \Delta m_{t-p})$ in (8b) with the equivalent set of regressors $(\Delta y_t, \Delta m_{t-1}, \Delta^2 y_t, \Delta^2 y_{t-1}, \ldots, \Delta^2 y_{t-p+1}, \Delta^2 m_{t-1}, \ldots, \Delta^2 m_{t-p+1})$. In (A1), the long-run multiplier is $\gamma_{my} = \alpha_{my}(1)/(1 - \beta_{mm})$, so that $\alpha_{my}(1) = \gamma_{my} - \beta_{mm}\gamma_{my}$. Making this substitution, (A1) can be written as

$$\Delta m_t - \gamma_{my}\Delta y_t = \beta_{mm}(\Delta m_{t-1} - \gamma_{my}\Delta y_t) + \sum_{j=0}^{p-1} \tilde{\alpha}_{my}^{j}\Delta^2 y_{t-j} + \sum_{j=0}^{p-1} \tilde{\alpha}_{mm}^{j}\Delta^2 m_{t-j} + \epsilon_t^m. \tag{A2}$$

Equation (A2) can be estimated by instrumental variables by regressing $\Delta m_t - \gamma_{my}\Delta y_t$ onto $(\Delta m_{t-1} - \gamma_{my}\Delta y_t, \Delta^2 y_t, \Delta^2 y_{t-1}, \ldots, \Delta^2 y_{t-p+1}, \Delta^2 m_{t-1}, \ldots, \Delta^2 m_{t-p+1})$ using $\{\Delta y_{t-i}, \Delta m_{t-i}\}_{i=1}^{p}$ as instruments. (Instruments are required because of the potential correlation between Δy_t and the error term.) Equation (8a) can now be estimated by instrumental variables using the residual from the estimated (A2) together with $\{\Delta y_{t-i}, \Delta m_{t-i}\}_{i=1}^{p}$. When a value for γ_{ym} is used to identify the model, this process was reversed.

Two complications arise in the calculation of standard errors for the estimated models. The first is that the long-run multipliers, γ_{ym} and γ_{my}, are

nonlinear functions of the regression coefficients. Their standard errors are calculated from standard formula derived from delta method arguments. The second complication arises because one of the equations is estimated using instruments that are residuals from another equation. This introduces the kind of "generated regressor" problems discussed in Pagan (1984). To see the problem in our context, notice that all of the models under consideration can be written as

$$y_t^1 = x_t^{1'}\delta_1 + \epsilon_t^1 \tag{A3}$$

$$y_t^2 = x_t^{2'}\delta_2 + \epsilon_t^2. \tag{A4}$$

Where, for example, when λ_{my} is assumed known, $y_t^1 = \Delta m_t - \lambda_{my}\Delta y_t, x_t^1$ represents the set of regressors $\{\Delta y_{t-i}, \Delta m_{t-i}\}_{i=1}^p, y_t^2 = \Delta y_t$, and x_t^2 represents the set of regressors $[\Delta m_t, \{\Delta y_{t-i}, \Delta m_{t-i}\}_{i=1}^p]$. Alternatively, when γ_{my} is assumed known, $y_t^1 = \Delta m_t - \gamma_{my}\Delta y_t, x_t^1$ represents the set of regressors $[\Delta m_{t-1} - \gamma_{my}\Delta y_t, \Delta^2 y_t, \{\Delta^2 y_{t-i}, \Delta^2 m_{t-i}\}_{i=1}^{p-1}]$, $y_t^2 = \Delta y_t$, and x_t^2 represents the set of regressors $[\Delta m_t, \{\Delta y_{t-i}, \Delta m_{t-i}\}_{i=1}^p]$.

Equations (A3) and (A4) allow us to discuss estimation of all the models in a unified way. First, (A3) is estimated using $z_t = \{\Delta y_{t-i}, \Delta m_{t-i}\}_{i=1}^p$ as instruments. Next, equation (A4) is estimated using $\hat{u}_t = (\hat{\epsilon}_t^1, z_t')$ as instruments, where $\hat{\epsilon}_t^1$ is the estimated residuals from (A3). If ϵ_t^1 rather than $\hat{\epsilon}_t^1$ is used as an instrument, standard errors could be calculated using standard formulae. However, when $\hat{\epsilon}_t^1$, an estimate of ϵ_t^1, is used, a potential problem arises. This problem will only effect the estimates in (A4) since $\hat{\epsilon}_t^1$ is not used as an instrument in (A3).

To explain the problem, some additional notation will prove helpful. Stack the observations for each equation so that the model can be written as

$$Y_1 = X_1\delta_1 + \epsilon_1 \tag{A5}$$

$$Y_2 = X_2\delta_2 + \epsilon_2, \tag{A6}$$

where Y_1 is $T \times 1$, etc. Denote the matrix of instruments for the first equation by Z, the matrix of instruments for the second equation by $\hat{U} = [\hat{\epsilon}_1\ Z]$, and let $U = [\epsilon_1\ Z]$. Since $\hat{\epsilon}_1 = \epsilon_1 - X_1(\hat{\delta}_1 - \delta_1), \hat{U} = U - [X_1(\hat{\delta}_1 - \delta_1)\ 0]$. Let $V_1 = \sigma_{\epsilon_1}^2$ plim $[T(Z'X_1)^{-1}(Z'Z)(X_1'Z)]$ denote the asymptotic covariance matrix of $T^{1/2}(\hat{\delta}_1 - \delta_1)$.

Now write,

$$\begin{aligned} T^{1/2}(\hat{\delta}_2 - \delta_2) &= (T^{-1}\hat{U}'X_2)^{-1}(T^{-1/2}\hat{U}'\epsilon_2) = (T^{-1}\hat{U}'X_2)^{-1}(T^{-1/2}U'\epsilon_2) \\ &\quad -(T^{-1}\hat{U}'X_2)^{-1}\begin{bmatrix} T^{1/2}(\hat{\delta}_1 - \delta_1)'(T^{-1}X_1'\epsilon_2) \\ 0 \end{bmatrix}. \end{aligned} \tag{A7}$$

It is straightforward to verify that plim $T^{-1}\hat{U}'\hat{U}$ = plim $T^{-1}U'U$ and that $T^{-1}\hat{U}'X_2$ = plim $T^{-1}U'X_2$. Thus, the first term on the right-hand side of (A7)

is standard: it is asymptotically equivalent to the expression for $T^{1/2}(\hat{\delta}_2 - \delta_2)$ that would obtain if U rather than $\hat{U}$ were used as instruments. This expression converges in distribution to a random variable distributed as $N(0, \sigma^2_{\epsilon_2}$ plim $[T(\hat{U}'X_2)^{-1}(\hat{U}'\hat{U})(X_2'\hat{U})^{-1}])$, which is the usual expression for the asymptotic distribution of the IV estimator.

Potential problems arise because of the second term on the right-hand side of (A7). Since $T^{1/2}(\hat{\delta}_1 - \delta_1)$ converges in distribution, the second term can only be disregarded asymptotically when plim $T^{-1}X_1'\epsilon_2 = 0$, that is, when the regressors in (A3) are uncorrelated with the error terms in (A4). In our context, this will occur when λ_{my} and λ_{ym} are assumed known, since in this case x_t^1 contains only lagged variables. However, when γ_{my} or γ_{ym} are assumed known, x_t^1 will contain the contemporaneous value of Δy_t or Δm_t, and thus x_t^1 and ϵ_t^2 will be correlated. In this case the covariance matrix of $\hat{\delta}_2$ must be modified to account for the second term on the right-hand side of (A7).

The necessary modification is as follows. Standard calculations show that $T^{1/2}(\hat{\delta}_1 - \delta_1)$ and $T^{-1/2}U'\epsilon_2$ are asymptotically independent under the maintained assumption that $E(\epsilon_2|\epsilon_1) = 0$; thus, the two terms on the right-hand side of (A7) are asymptotically uncorrelated. A straightforward calculation demonstrates that $T^{1/2}(\hat{\delta}_2 - \delta_2)$ converges to a random variable with a $N(0, V_2)$ distribution where

$$V_2 = \sigma^2_{\epsilon_2} \text{ plim } [T(\hat{U}'X_2)^{-1}(\hat{U}'\hat{U})(X_2'\hat{U})^{-1}] \; + \; \text{plim } [T(\hat{U}'X_2)^{-1}D(X_2'\hat{U})^{-1}],$$

where D is a matrix with all elements equal to zero, except that $D_{11} = (\epsilon_2'X_1)TV_1(X_1'\epsilon_2)$, and where $TV_1 = \sigma^2_{\epsilon_1}(Z'X_1)^{-1}(Z'Z)(X_1'Z)^{-1}$. Similarly, it is straightforward to show that the asymptotic covariance between $T^{1/2}(\hat{\delta}_1 - \delta_1)$ and $T^{1/2}(\hat{\delta}_2 - \delta_2) = -\text{plim}[V_1(T^{-1}X_1'\epsilon_2)\ 0][T^{-1}X_2'\hat{U}]$.

An alternative to this approach is the GMM-estimator in Hausman, Newey, and Taylor (1987). This approach considers the estimation problem as a GMM problem with moment conditions $E(z_t\epsilon_t^1) = 0, E(z_t\epsilon_t^2) = 0$, and $E(\epsilon_t^1\epsilon_t^2) = 0$. The GMM approach is more general than the one we have employed, and when the errors terms are non-normal and the model is over-identified, it may produce more efficient estimates.

REFERENCES

Aschauer, David, and Jeremy Greenwood. "A Further Exploration in the Theory of Exchange Rate Regimes," *Journal of Political Economy,* vol. 91 (October 1983), pp. 868–72.

Barro, Robert J., and Mark Rush. "Unanticipated Money and Economic Activity," in Stanley Fischer, ed., *Rational Expectations and Economic Policy*. Chicago: University of Chicago Press, 1980.

Bernanke, Ben S. "Alternative Explanations of the Money-Income Correlation," *Carnegie-Rochester Conference Series on Public Policy*, vol. 25 (Autumn 1986), pp. 49–99.

Beveridge, Stephen, and Charles R. Nelson. "A New Approach to Decomposition of Economic Time Series into Permanent and Transitory Components with Particular Attention to Measurement of the 'Business Cycle,'" *Journal of Monetary Economics*, vol. 7 (March 1981), pp. 151–74.

Blanchard, Olivier J. "A Traditional Interpretation of Macroeconomic Fluctuations," *American Economic Review*, vol. 79 (December 1989), pp. 1146–64.

_____, and Danny Quah. "The Dynamic Effects of Aggregate Demand and Supply Disturbances," *American Economic Review*, vol. 79 (September 1989), pp. 655–73.

_____, and Mark Watson. "Are Business Cycles All Alike?" in Robert J. Gordon, ed., *The American Business Cycle: Continuity and Change*. Chicago: University of Chicago Press, 1986.

Christiano, Lawrence, and Martin Eichenbaum. "Liquidity Effects, Monetary Policy, and the Business Cycle," *Journal of Money, Credit, and Banking*, vol. 27 (November 1995), pp. 113–36.

Cooley, Thomas F., and Gary D. Hansen. "The Inflation Tax in a Real Business Cycle Model," *American Economic Review*, vol. 79 (September 1989), pp. 733–48.

Economic Report of the President, 1969. Washington: Government Printing Office, 1969.

Evans, Martin D. D., and Karen L. Lewis. "Do Expected Shifts in Inflation Affect Estimates of the Long-Run Fisher Relation?" Manuscript. University of Pennsylvania, 1993.

Fisher, Mark E., and John J. Seater. "Long-Run Neutrality and Superneutrality in an ARIMA Framework," *American Economic Review*, vol. 83 (June 1993), pp. 402–15.

Fuerst, Timothy S. "Liquidity, Loanable Funds, and Real Activity," *Journal of Monetary Economics*, vol. 29 (February 1992), pp. 3–24.

Gali, Jordi. "How Well Does the IS-LM Model Fit Postwar U.S. Data?" *Quarterly Journal of Economics*, vol. 107 (May 1992), pp. 709–38.

Geweke, John. "The Superneutrality of Money in the United States: An Interpretation of the Evidence," *Econometrica*, vol. 54 (January 1986), pp. 1–21.

Goldberger, Arthur S. *Econometric Theory*. New York: John Wiley and Sons, 1964.

Gordon, Robert J. "The Recent Acceleration of Inflation and Its Lessons for the Future," *Brookings Papers on Economic Activity,* 1:1970, pp. 8–41.

Hausman, Jerry A., Whitney K. Newey, and William E. Taylor. "Efficient Estimation and Identification of Simultaneous Equation Models with Covariance Restrictions," *Econometrica,* vol. 55 (July 1987), pp. 849–74.

Johnston, J. *Econometric Methods,* 3d ed. New York: McGraw Hill, 1984.

King, Robert G., and Charles I. Plosser. "Money Business Cycles," *Journal of Monetary Economics,* vol. 33 (April 1994), pp. 405–38.

_____, Charles I. Plosser, James H. Stock, and Mark W. Watson. "Stochastic Trends and Economic Fluctuations," *American Economic Review,* vol. 81 (September 1991), pp. 819–40.

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

_____. "Testing Long-Run Neutrality," Working Paper 4156. Boston: National Bureau of Economic Research, September 1992.

Lucas, Robert E., Jr. "Liquidity and Interest Rates," *Journal of Economic Theory,* vol. 50 (April 1990), pp. 237–64.

_____. "Some International Evidence on Output-Inflation Trade-offs," *American Economic Review,* vol. 63 (June 1973), pp. 326–34.

_____. "Econometric Testing of the Natural Rate Hypothesis," in Otto Eckstein, ed., *The Econometrics of Price Determination.* Washington: Board of Governors of the Federal Reserve System, 1972.

McCallum, Bennett T. *Monetary Economics: Theory and Policy*. New York: Macmillan, 1989.

_____. "On Low-Frequency Estimates of Long-Run Relationships in Macroeconomics," *Journal of Monetary Economics,* vol. 14 (July 1984), pp. 3–14.

Mehra, Yash P. "Some Key Empirical Determinants of Short-Term Nominal Interest Rates," Federal Reserve Bank of Richmond *Economic Quarterly,* vol. 81 (Summer 1995), pp. 33–51.

Mishkin, Frederic S. "Is the Fisher Effect Real? A Reexamination of the Relationship between Inflation and Interest Rates." Manuscript. Columbia University, 1992.

Pagan, Adrian. "Econometric Issues in the Analysis of Regressions with Generated Regressors," *International Economic Review,* vol. 25 (February 1984), pp. 221–48.

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

Rotemberg, Julio J., John C. Driscoll, and James M. Poterba. "Money, Output, and Prices: Evidence from a New Monetary Aggregate," *Journal of Economic and Business Statistics,* vol. 13 (January 1995), pp. 67–84.

Sargent, Thomas J. "A Classical Macroeconometric Model for the United States," *Journal of Political Economy,* vol. 84 (April 1976), pp. 207–37.

_____. "A Note on the Accelerationist Controversy," *Journal of Money, Credit, and Banking,* vol. 3 (August 1971), pp. 50–60.

Shapiro, Matthew, and Mark W. Watson. "Sources of Business Cycle Fluctuations," National Bureau of Economic Research *Macroeconomics Annual,* vol. 3 (1988), pp. 111–56.

Sims, Christopher A. "Models and Their Uses," *American Journal of Agricultural Economics,* vol. 71 (May 1989), pp. 489–94.

_____. "Are Forecasting Models Usable for Policy Analysis?" Federal Reserve Bank of Minneapolis *Quarterly Review,* vol. 10 (Winter 1986), pp. 2–16.

Solow, Robert. *Price Expectations and the Behavior of the Price Level.* Manchester, U.K.: Manchester University Press, 1969.

Stock, James H. "Confidence Intervals for the Largest Autoregressive Root in U.S. Macroeconomic Time Series," *Journal of Monetary Economics,* vol. 28 (December 1991), pp. 435–60.

_____, and Mark W. Watson. "Interpreting the Evidence on Money-Income Causality," *Journal of Econometrics,* vol. 40 (January 1989), pp. 161–81.

Stockman, Alan C. "Anticipated Inflation and the Capital Stock in a Cash-In-Advance Economy," *Journal of Monetary Economics,* vol. 8 (November 1981), pp. 387–93.

Tobin, James. "Money and Economic Growth," *Econometrica,* vol. 33 (October 1965), pp. 671–84.

Watson, Mark W. "Vector Autoregressions and Cointegration," in Robert Engle and Daniel McFadden, eds., *Handbook of Econometrics,* Vol. IV. Amsterdam: Elsevier, 1994.