Are On-Line Currencies Virtual Banknotes?

STEPHEN F. QUINN AND WILLIAM ROBERDS

Quinn is an associate professor of economics at Texas Christian University. Roberds is a vice president and economist in the Atlanta Fed's research department.
They thank the many people who commented on previous drafts, especially James McAndrews and Lawrence White, who served as reviewers. Although the discussion in this article touches on certain legal issues, the authors are laypersons in the law, and their remarks should not be construed as legal advice.

Cash is increasingly being displaced by private forms of payment. Currently the U.S. economy functions with a minimal stock of cash, probably amounting to less than 2.6 percent of its annual gross domestic product (GDP).[1] This figure is markedly less than historical estimates for the United States (for example, about 3.2 percent in 1960) or contemporary estimates for other countries (as high as 4.9 percent for some European countries, according to Humphrey 2002). Roughly three-quarters of all transactions still take place on a cash basis (Committee on the Federal Reserve in the Payments Mechanism 1998), but the average amount of a cash-based transaction is small, probably less than $10.[2] When payment technologies are compared on a value basis, payments based on the transfer of "inside money" (payments by check, payment card, or direct transfer) dominate, accounting for the vast majority of the value of transactions within the United States.[3]

Payment in inside money is, of course, hardly a recent phenomenon. By the fourteenth century, European merchants had discovered the essential advantage of inside money: Exchange using debt ties up fewer resources than does the exchange of costly coin.[4] Since not everyone's debt is likely to be equally reliable, however, inside-money payment systems have historically singled out the debt of a select group of "strong credits" (banks) as closer proxies for commodity (or outside) money. These privileged forms of debt possess the moneylike property of finality—of being able to extinguish other debts by virtue of their transfer from debtor to creditor.[5] However, the limitation of this privilege to certain strong credits also imposes constraints on those parties whose debt does not qualify as money. Hence, there has been an incentive to extend the reach of inside money with payment devices of limited finality, such as the check. Such instruments can broaden the benefits of inside money but may also increase the risk of default or fraud.

Monetary history is punctuated by innovations—deposit banking, checks, banknotes, credit cards—that have expanded the role of inside money. For example, in recent years technology has made it possible for virtually anyone with a credit or debit card to pay for any purchase (from a merchant with an account with a credit card company) anywhere with a relatively high degree of finality. In many situations, card-based payment systems have offered considerable improvements over their paper-based predecessors.[6] A merchant selling a good to an unfamiliar customer can accept a card payment with the confidence that such payment is usually, if not completely, final.[7] Payment by check would not offer the merchant the same degree of finality, and requiring cash payment could deny customers access to credit.

The finality associated with card payments does not extend to every transaction environment, however.

Payment cards, and especially credit cards, are often used in situations—such as mail order, telephone, and Internet transactions—in which the cardholder is not present and cannot sign a receipt. In such cases the risk of fraud is elevated, but little of this risk is borne by credit card holders because (under U.S. law at least) their liability is limited to $50 and in practice is often zero.[8] A credit card holder may also withhold payment if he believes he has been charged for goods or services that were not delivered or were defective. In such circumstances, offering blanket guarantees of payment finality to merchants would create an unmanageable risk for card issuers. Instead, merchants bear most of the fraud risk in the form of liability for chargebacks (debits to a merchant's account resulting from disputed payments) from the card issuers. This risk allocation has made "cardholder not present" credit card payment more expensive and generally less attractive to merchants unwilling to accept the risk of chargebacks. Internet transactions seem especially at risk, and this riskiness is reflected in fraud rates for on-line transactions. Trade publications have reported rates of credit card fraud as high as 2.1 percent for Web-based transactions, roughly ten times the rate for face-to-face transactions.[9]

Despite the obvious differences between on-line currencies and physical banknotes, they share some conspicuous similarities in the circumstances of their birth.

The past few years have seen the debut of several new types of on-line payment arrangements, at least partly in response to the difficulties associated with card-based payment over the Internet. These arrangements offer the promise of making it possible for anyone to pay anyone on-line, even in situations in which card-based payment would be infeasible or uneconomical. The most innovative arrangements, sometimes referred to as on-line currencies, bypass the traditional, bank-based methods for clearing and settlement of payments in favor of a simple "on-us" funds transfer—that is, a transfer of a claim on the on-line currency issuer (in the form of an account balance) from payor to payee.[10] While the finality of such transfers has thus far been of a limited nature, the most successful on-line currency issuer, PayPal, now offers its users finality guarantees under some circumstances.

What is the future of this type of payment arrangement? To date, industry reviews have been mixed. Most observers concede that on-line currencies have offered a useful service for person-to-person on-line transactions, most typically those associated with on-line auctions. On the other hand, on-line currencies have seen relatively little use in purchases by consumers from businesses, and most of these exchanges have been restricted to small enterprises. This situation has led some analysts to believe that future use of on-line currencies will be, at best, restricted to the person-to-person niche.

This article examines the likely success or failure of on-line currencies by means of a historical analogy. Specifically, the discussion compares the introduction of on-line currencies to the debut of the bearer banknote, the direct predecessor to modern currency, in late-seventeenth-century London. Despite the obvious differences between these on-line currencies and everyday, physical banknotes, the argument presented here will show that they share some conspicuous similarities in the circumstances of their birth. In particular, the article argues that the key innovation of the earliest banknotes was to provide finality under circumstances in which extant payment systems either could not ensure final payment or could do so only at an unacceptable cost. The next section describes how on-line currencies may be able to fill the same role in the context of e-commerce. The discussion concludes with some observations about future prospects for on-line currencies, again using the (clearly successful) introduction of the banknote as a historical model.

Early Forms of Inside-Money Payment

An initial summary of the prebanknote payment system in Europe, which combined deposit banking, orders to transfer deposits, and transfer of those orders by endorsement, is helpful in explaining the innovation offered by banknotes and the potential for on-line currencies. The system began with deposit banking in Italy, where two merchants desiring to transfer funds would together visit a banker and have one account debited and the other credited. Such transfers *in banco* spared merchants the transportation, protection, assay, and opportunity costs of using coin—the outside money of the time. The banker's ledger formed a permanent record, and payment within the bank was final.

To avoid the need for both parties to visit the bank together, deposit banking developed payment by check or draft. Checks drawn on banks in early

modern Europe, including the goldsmith bankers of seventeenth-century London, fulfilled a role similar to that of personal checks drawn on modern deposit banks—such checks enhanced decentralized exchange. Then, as now, the convenience of payment by check created a risk of default because payment was not final until the bank honored the check, and then, as now, whether the bank honored the check depended on the adequacy of the check drawer's account balance or the willingness of the bank to allow an overdraft. This risk was manageable, but only because checks were generally used by prominent personages and for local payments only.

To arrange the payment of funds outside the local banking system, one had to arrange for payment by bill of exchange. Much like a modern traveler's check, a bill ordered someone in a distant location to pay a fixed sum to a payee at that location. However, a bill was different from a modern traveler's check in that it was payable only after some fixed amount of time had passed. Bills of exchange were generally payable in the prevalent currency of the distant location. For a bill to work, the person who wrote the bill (the drawer) had to arrange for someone to pay the bill at the other end (the acceptor). This arrangement was most easily made if the drawer had a close relationship with the acceptor. For example, Renaissance Italians established international family networks to act as acceptors. Later, bankers used systems of agents or correspondent banks. Once the bill had been accepted (always indicated in writing on the bill), it became a legally enforceable claim against the acceptor. Or the acceptor could refuse the bill by protesting it (and indicating so in writing on the bill) and returning it to the drawer.

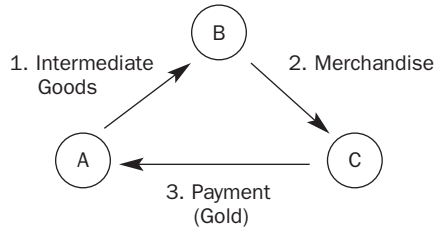
The transfer of checks, drafts, and bills of exchange extended the opportunity to use inside money beyond the immediate range of a deposit bank. Remote transfer of third-party debt had a beneficial netting effect, reducing a chain of obligations to a single obligation between the original obligor and the ultimate creditor. The benefits of remote transfer were especially pronounced for places that outlawed deposit banking, such as London and Antwerp (van der Wee 1997). Instead of checks and ledger entries, inside money in these locales had to take the form of circulating personal obligations.[11] A key advance in promoting extensive use of remote transfer was recognition of the legal standing of parties who had been assigned the debt

1. As of this writing, U.S. GDP is about $10.7 trillion. The stock of U.S. currency is about $690 billion, but, according to estimates (see Porter and Judson 1996), at least 60 percent of this stock resides outside the United States. The 2.6 percent figure is thus calculated as 40 percent of the currency/GDP ratio. Humphrey (2002), applying similar methods, arrives at an estimate of 1.7 percent.
2. Boeschoten (1992) estimates that the average value of a cash transaction in the United States is about $5.
3. "Inside money" is the term used by economists to refer to money created by the private sector, typically money in bank accounts. "Outside money" refers to money created outside the private sector, meaning currency issued by, or money held in accounts at, a central bank. This generalization about inside money holds even when large-value, interbank settlements are excluded. The Bank for International Settlements (2001) estimates that daily U.S. noncash transactions averaged $288 billion in 1999, not counting interbank settlements. If the average cash transaction amounted to, say, $20, this scenario would imply that the value of cash transactions makes up only 5 percent of the value of all transactions.
4. On the early use of inside money, see, for example, De Roover (1948) for Bruges, Usher (1943) for Barcelona, and Mueller (1997) for Venice. Today inside money supplants outside (government-issued) money instead of metallic coin.
5. Finality was a key feature of early banking arrangements. For example, De Roover (1948, 335) observes that oral transfers of bank deposits were irreversible once a transfer had been recorded in the bank's ledger.
6. This improvement stems from the fact that credit cards have reduced both payment risks and, in many cases, costs (see Berger, Hancock, and Marquardt 1996, 700–709).
7. Credit card companies often absorb the loss in cases in which the merchant has obtained authorization from the credit card company for the transaction and has also obtained the customer's signature. A recent study by the U.S. General Accounting Office (1997, 114) reports that the card companies' average share of losses on credit card transactions is 70 percent, with the remainder borne by the merchants.
8. In the case of credit cards, a cardholder's liability in cases of fraud is limited by the Truth in Lending Act of 1968 (TILA). TILA also guarantees cardholders the right to withhold payment in certain instances. See, for example, Mann (1999, 107–40) for a detailed discussion of TILA and its implications.
9. See, for example, Punch (2002) or Lee (2003). The fraud rate is typically calculated as the value of fraudulent transactions as a fraction of the value of all transactions.
10. Many other names have been proposed for these arrangements. Two of the most common are "alternative currency" and "online payment systems." Kuttner and McAndrews (2001) employ the term "proprietary account systems" while Schreft (2002) uses "proprietary monetary value." For the limited purposes of this article, all of these terms will be considered synonymous.
11. These obligations were either bills of exchange or personal promises to pay, then called letters obligatory and later called promissory notes.

FIGURE 1

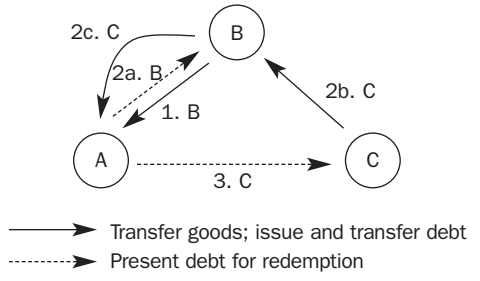
A Bill-of-Exchange Transaction

A: Flow of Goods



Sequence of Events
1. Supplier A transfers intermediate goods to merchant B.
2. Merchant B transfers merchandise to customer C (in return for bill).
3. Customer C honors bill and pays supplier A.

B: Flow of Debt



Sequence of Events
1. Merchant B issues debt to supplier A.
2a. Supplier A presents B's debt for redemption.
2b. Customer C issues debt (bill) to merchant B.
2c. Merchant B transfers C's debt to supplier A.
3. Supplier A presents C's debt for redemption.

of a third party in payment. The London Mayor's Court granted such recognition in 1436, and the concept spread to Antwerp (Munro 2000).

Even with legal recognition, the effectiveness of remote transfers without banks was limited because information was needed to assess the credibility of the debt issuer (the acceptor of a bill), and such information was often asymmetric and idiosyncratic. Transfer created an incentive to pass on high-risk or fraudulent debt. In 1507, Antwerp mitigated this problem by creating a legal obligation of contingent liability on anyone who transferred third-party debt (van der Wee 1997, 325). According to the new rule, when a payor paid in the debt of a third party, the payor was also obligated to accept liability for the debt should the original obligor (or previous transferors of the debt) be unable to settle. Contingent liability gave anyone who wanted to circulate debt a strong incentive to screen the quality of the debt he was attempting to circulate. In practice, the simplest way of recording who had transferred a debt was to have each party sign the back of the debt.[12] The institution of endorsement (transfer with contingent liability by means of a signature) spread across Europe and was applied to checks and bills of exchange. Combining legal standing with transfer by endorsement gave rise to the concept of a negotiable instrument, essentially a freely transferable debt whose possession automatically confers upon its holder well-understood rights as a creditor.[13] Amsterdam became the dominant hub of international finance by buttressing a payment system based on the exchange of negotiable instruments with a municipal exchange bank (Dehing and 't Hart 1997).

Another distinctive feature of negotiable instruments was the idea that anyone receiving an instrument by means of endorsement became a "holder in due course."[14] Essentially this concept meant that endorsees almost always enjoyed full creditor's rights, even in cases when the good that was supposed to be delivered against the original obligation was not delivered or was defective (with some exceptions for sham transactions associated with fraud schemes). This feature enhanced the "moneyness" of negotiable debt by ensuring that good-faith transfers of such debt were final, barring default of the original obligor.

A Model of Debt Transfer

Kahn and Roberds (2001) analyze debt transfer and circulation by endorsement in a formal economic model in which payment by transfer of negotiable debt results in a desirable allocation of risks among payor, payee, and outside parties. They consider a stylized example in which party A supplies an intermediate good to merchant B, who uses the intermediate good to produce a durable final good, merchandise. Merchandise is delivered to customer C in return for a promise of future payment (see Figure 1). However, C may default on the promised repayment for one of several reasons (C may change his mind about the value he places on the merchandise or may be subject to an event such as fraud). Of course, knowledge of his own propensity to change his mind is C's private information. Knowledge of the customer's susceptibility to fraud risk is also private information, but the merchant may have some better knowledge of this informa-

tion than the supplier does. All contracts between parties are subject to limited enforcement in the sense that assets held by a party defaulting on an obligation are not always attachable by creditors.

Optimal payment arrangements in this environment have two salient features. First, overly risky customers (those who have decided they do not want the merchandise or those too susceptible to credit events) should not receive merchandise. Second, in cases in which the merchandise is delivered, some portion of the promised payments by the customer should flow directly from the customer to the supplier, bypassing the merchant. In the latter case, an optimal allocation of risks can be implemented by a pair of debt contracts, one from the customer to the merchant and the other from the merchant to the supplier, as long as the merchant can discharge his debt by transferring the customer's debt to the supplier (see Figure 1B). In other words, the merchant uses the customer's debt to pay his own.

A potential problem with this type of arrangement is "adverse selection." That is, in cases when the merchant deals directly with the customer and the supplier does not, the merchant is apt to have better information about the customer's creditworthiness than is the supplier. The merchant may then have an incentive to pass on the debt of less creditworthy or nonexistent customers to the supplier. To guard against this temptation, the merchant must accept contingent liability for (endorse) the customer's debt should the customer be unable or unwilling to pay. For this endorsement to be meaningful, the merchant himself must have sufficient wealth at stake.

The intuition behind this result is straightforward. Payment by transfer of debt is desirable because it short-circuits the credit chain from customer to merchant to supplier, thereby limiting the possibilities for successive defaults. Transfer, however, creates an adverse selection problem, so adding endorsement gives the merchant an incentive to avoid transactions with overly risky customers.

Enter Banknotes

The combination of local deposit banking and circulating debt via endorsement created a successful system of inside money for the commercial elite but left out many people. Merchants, nobles, and others with sufficient standing could pay local obligations by means of checks drawn on a local bank, but these checks were useless for trading at a distance. Prominent firms could pay obligations incurred in long-distance trade by drawing bills payable on their overseas branches, but this option was out of the question for smaller firms. Likewise, large players could introduce others' bills into circulation by endorsing them over to their creditors, but such players had to have sufficient wealth (and sufficient information regarding the creditworthiness of the acceptor) to have their endorsements valued.

The key innovation of the earliest banknotes was to provide finality under circumstances in which extant payment systems either could not ensure final payment or could do so only at an unacceptable cost.

Mengle (1990) describes payment by check (and, by extension, similar negotiable instruments) as enforcing a loss-allocation rule that obeys a least-cost avoider principle. By requiring an endorsement with every transfer, this rule assigns liability for credit risk and fraud to the party presumed able to avoid such risks at least cost—the endorser. Mengle notes that for this type of rule to be effective, the party in question (in this case, the endorser) must be able both to bear the relevant risk and to undertake actions that contain the risk. Early users of negotiable instruments understood these limitations and restricted the use of such instruments accordingly.

To serve those excluded from the endorsement system, bankers in mid-seventeenth-century London developed a niche product that became the banknote.[15] Deposit banks finally appeared in London with the loosening of economic controls by Oliver Cromwell in the 1650s. Unlike in Amsterdam,

12. Originally, endorsements were always made to a specific party. There were no "endorsements in blank" such as those commonly entered on the back of a modern check.
13. The technical definition of negotiability is somewhat involved. For a discussion, see, for example, Winn (1998). Negotiability, while largely irrelevant for electronic payments, remains the basis for U.S. law pertaining to checks.
14. See, for example, Winn (1998) for a discussion. The notion of a holder in due course exists even in contemporary payment law but is of limited relevance in most situations.
15. Banknotelike instruments had seen sporadic use before this time. DeRosa (2001), for example, documents the issue of banknotes by Neapolitan public banks in the sixteenth century, although these notes generally circulated only by endorsement. The focus here is on the early London banknotes as the most direct predecessors of the modern, bearer banknote.

private banks in London were free to develop, and they rapidly did by offering services such as deposit accounts, money changing, lending, discounting, and international payments (Richards 1929, 23–24). The system was tied together by mutual acceptance with bilateral clearing, and some bankers even became key government financiers and tax collectors (Quinn 1997). Under this system, a customer who lacked enough personal renown to write a check could pay with a draft drawn by a banker on himself (somewhat analogous to a modern cashier's check). The practice of banks issuing drafts was a small step from the personal pledges common to English commercial practice.

Combining legal standing with transfer by endorsement gave rise to the concept of a negotiable instrument, essentially a freely transferable debt whose possession automatically confers upon its holder well-understood rights as a creditor.

The banker's note could then circulate by endorsement, but a signature added little value because people did not need to screen the banker within the local payment system. Since information about default was symmetrical (between payor and payee), the endorsement neither revealed information nor reduced any moral hazard. The benefit of an endorsement to the endorsee was essentially the same as the cost of the contingent liability to the endorser, so the situation had no least-cost avoider. Moreover, if carrying out the legal claims created by the contingent liability created unrecoverable costs such as time and legal fees, then endorsement could even become undesirable. The alternative was to allow for transfer without endorsement by making the banker's note payable to bearer. The combination of a banker's draft with the payable-to-bearer feature effectively achieved finality without having to wait for a draft to return for settlement by the issuing banker. Customers who needed bank-issued debt for payment purposes could easily prefer the bearer form, and this form of transfer created no additional cost for the banker.[16]

In the context of the Kahn-Roberds model, issue of one of these early banknotes can be interpreted as a sort of debt swap (see Figure 2). The earlier example is augmented by the addition of an agent known as the banker, who is informed with respect to the risk of party C. The banker has no access to productive technology but has an income that is verifiable by all parties. Merchant B swaps the debt of C, seen by B's supplier (A) as too risky, for the bearer debt issued by banker E. All parties gain from this transaction: A views E's debt as trustworthy and is therefore willing to supply B with an intermediate good in return for it. C is able to obtain merchandise in return for an uncertain promise of future repayment, and E is able to profit from his knowledge of C's creditworthiness and from his own verifiable wealth.[17]

A key characteristic of this arrangement, vis-à-vis earlier arrangements, was the transfer of risks in the payment process to the banker who issued the bearer note (though the other parties still bore the risk that the banker's note could be counterfeit). Specifically, if a bearer note was issued against a bill that was fraudulent or simply not repaid, the banker could have borne the loss. This risk allocation could again be described as obeying a least-cost avoider principle, only with the role of the least-cost avoider played by an outside party (the banker, party E) and not by a principal in the chain of transactions (an endorser, party B).

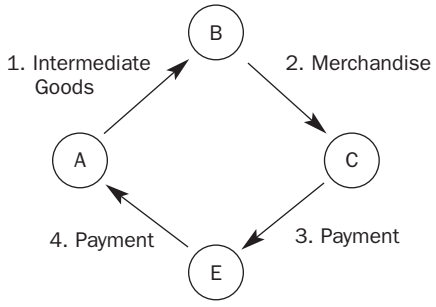
Bearer notes especially suited customers who were not well known and who did not have a local bank account but who did have an asset to offer. An example would be an unknown foreign merchant (or his agent) who had a bill drawn on someone most London merchants did not know. If the banker had an informational advantage because of his wide network of dealings, then the banker could buy the bill at a discount in exchange for a bearer banknote. The merchant received a local means of payment with finality, and the banker profited from his expertise. Bearer banknotes were a financial innovation that extended the immediacy of settlement beyond walking distance from a banker's ledger.

The earliest extant ledgers of a London banker, those of Edward Backwell, confirm the modest beginnings of banknotes. Backwell was a member of the first cohort of goldsmiths to open deposit banks in London sometime in the 1650s. According to his earliest surviving ledger, in 1663 Backwell was already a full-service deposit banker. Backwell was a prominent banker, one who was even mentioned in Samuel Pepys's famous diary, because he invested heavily in government debt and managed the tax farm that collected the customs. He was ruined by the government's default in 1672. His records from 1663 through 1671 survived because his heirs married into the Childs banking family, whose bank of the same name still operates on Fleet Street.

FIGURE 2

A Banknote Transaction

A: Flow of Goods

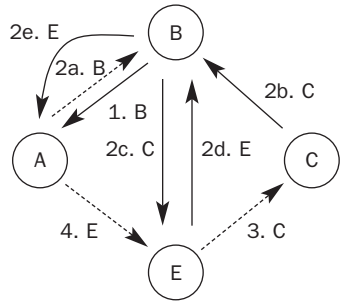


Sequence of Events

1. Supplier A transfers intermediate goods to merchant B.
2. Merchant B transfers merchandise to customer C in return for C's debt, which is transferred to banker E.
3. Customer C pays debt to banker E.
4. Banker E pays debt to supplier A.

Note: Events 3 and 4 may occur in reverse order.

B: Flow of Debt



Transfer goods; issue and transfer debt

Present debt for redemption

Sequence of Events

1. Merchant B issues debt to supplier A.
2a. Supplier A presents B's debt for redemption.
2b. Customer C issues debt (bill) to merchant B.
2c. Merchant B transfers C's debt to banker E.
2d. Banknote issued by banker E to merchant B.
2e. Merchant B discharges debt to supplier A by transferring banknote.
3. Banker E presents customer C's debt for redemption.
4. Supplier A presents E's banknote.

In 1668 Backwell began a small business issuing bearer notes separate from any demand account (Royal Bank of Scotland, *Backwell's Ledger Q,* folio 274). Depositors with Backwell may have been accepting bearer notes even earlier; unfortunately, the ledgers do not explain how depositors withdrew their funds. The "bearer" account, however, explicitly records the creation of bearer banknotes outside of a deposit account. From 1668 to 1671 he issued sixteen such bearer notes with an average value of £174, a tiny sum given the size of his bank. Over the same period, for example, Backwell paid customers a total of £75,000 in interest alone.[18] The banknotes circulated for only a few days. The exception was one experimental note, which circulated for ninety-two days, but this situation was made possible only by offering a 6.5 percent annual interest rate on the note. Available evidence thus suggests that the issuance of banknotes began as a side business. The profitability of the notes lay not in their circulation but most likely in their ability to "credit enhance" other forms of debt, that is, to be accepted in exchange for discounted debt drawn on other parties.

The banknote, however, was a scalable product. The prerequisites were recognized standing in the local community, an ability to assume the credit risk associated with discounting debt, and a well-informed position in the asset market. In 1694, just twenty-five years after Backwell's experimentation, the newly founded Bank of England was purchasing assets with banknotes on a massive scale. The Bank of England was a chartered joint-stock company with a subscription of £1.2 million. It had only a tiny

16. The bearer feature of such debt left no record of transactions. But the absence of such a record did not really place additional limits on the recourse of parties using such debt against seller-side fraud relative to the use of instruments that were payable to order. In the latter case, a seller would have been a holder in due course and fully entitled to enforce the debt against the original obligor (the banker).
17. The banker could also simply pay coin for C's debt. But to do so would entail an opportunity cost, namely, the cost of liquidating another asset in order to obtain the coin.
18. See Royal Bank of Scotland, *Backwell's Ledger*. For bearer notes, see *Ledger Q*, folio 274; *Ledger R*, folio 296; *Ledger S*, folio 335; and *Ledger T*, folio 83. For interest, see *Ledger Q*, folios 111, 421, 481, 521, 612, 621, 631–35; *Ledger R*, folios 121, 421, 599, 601–9; *Ledger S*, folios 141, 421, 582, 612, 630–38; and *Ledger T*, folios 61, 321, 451, 591, 612–21.

deposit business, yet by 1696 the Bank of England had issued £800,000 in banknotes (Horsefield 1983, 24). Unlike the earliest banknotes, which were most likely issued against commercial bills of exchange, the Bank of England's notes were used to acquire large amounts of government debt. The Bank of England did discount bills of exchange, but discounted bills were never a major asset on the Bank of England's balance sheet until the Napoleonic Wars, a century after the bank's founding.

The issue of such an outsized quantity of notes was essentially a solution to a chicken-and-egg problem. In contrast to earlier banknotes, Bank of England notes commonly remained in circulation even when they bore no interest. There was a network effect present in large-scale note issue in the sense that the liquidity of the notes (the willingness of counterparties to accept the notes as payment) grew as they became more widely held. The profitability of the notes grew as their liquidity increased because people were willing to hold the notes for their liquidity value and less likely to present them for redemption. Thus, a large-scale issue was needed for the notes' liquidity while liquidity was required to sustain large-scale issue. To benefit from these network effects, the Bank of England needed to execute a sizable initial swap of notes for debt. The government, with its large financing needs, provided an ideal counterparty for such a swap.[19]

The initial advantage of banknotes was their ability to provide payment finality in situations in which existing payment institutions could not efficiently do so.

By issuing bearer notes in such large quantity, the Bank of England also effected a qualitative change in the character of the notes. Rather than functioning as credit-enhanced versions of other, unmarketable obligations, banknotes were seen as general claims backed by the assets of the bank, which by and large consisted of government securities. Bearer notes were thus transformed from a niche product to a viable competitor for coin and bills as a medium of exchange.

The issue of banknotes became common practice in Anglo-American banking. Scotland established a system of corporate, note-issuing banks beginning with the Bank of Scotland in 1695 (Checkland 1975, 23–90; White 1984, 22–34). In the following century, discounting of bills of exchange with banknotes became the dominant means of finance for English banks outside of London (Pressnell 1956, 136–89). Banknote finance also dominated U.S. banks until the Civil War.[20]

Not all early, large-scale note issues were successful, and the issue of banknotes contributed directly to the famous twin financial debacles of 1720: the Mississippi Bubble in France and the South Sea Bubble in London. In both cases, banknotes were issued well beyond the value of government debt that the issuers actually held (Neal 1991, 62–117). The collapse of the Mississippi Bubble soured the French on note-issue banking for the remainder of the Old Regime. The collapse of the South Sea Company in London left the Bank of England the only corporate bank in England for over a century.[21]

Nor did bearer notes ever completely supplant coin, especially for small-value transactions. Banknotes were originally conceived as wholesale or business-to-business products, and this was their predominant use throughout their early history. Banknotes were typically issued in large denominations only. Small-denomination bearer notes were legally prohibited on the grounds that they were subject to moral hazard problems (Smith [1776] 1994, 351–52) or, worse, a potential hazard to the maintenance of a precious-metal standard.[22]

To summarize, bearer banknotes were an innovation in the payment system that began by serving a very small niche market. The initial advantage of banknotes was their ability to provide payment finality in situations in which existing payment institutions could not efficiently do so. The Bank of England was able to scale up this idea by pairing the public's liquidity demand with the government's considerable financing needs. The resulting arrangement went on to redefine the notion of money and to revolutionize government finance.

On-Line Payments

The Internet has accelerated the demand for inside money payments that do not involve face-to-face contact. Of the traditional payment technologies, credit cards have been the payment mode of choice for on-line transactions.[23] However, as noted earlier in this article, the finality of card-based payments over the Internet is rather less than in face-to-face transactions. While card issuers bear some of the credit risk associated with on-line transactions, on-line merchants bear most of the fraud risk.

The lack of finality in card-based on-line transactions places today's on-line merchants in a position somewhat comparable to that of merchants doing business in early modern Europe.[24] As was the case with the early merchants, today's on-line merchants often have little choice but to accept risky debt in payment. When a customer offers on-line payment via a credit card, the merchant can receive value for that payment only by negotiation and endorsement (in the sense of transferal with acceptance of conditional liability) of the customer's debt to other parties involved in the clearing of credit card payments.[25]

To place the on-line merchants' situation in the context of the model depicted in Figure 1, let B represent a merchant doing business on-line with a customer, C. Party A represents an amalgam of upstream parties with whom the merchant must deal in order to do business, including suppliers, merchant acquirers, card associations, and card issuers.[26] In taking an on-line card-based payment, the merchant is in effect taking the debt of the cardholder with the expectation that it can be transferred for value. Current rules concerning chargebacks limit the finality of such transfers, however, and require that the merchant be willing to accept liability for chargebacks. While the merchant is nominally protected from credit risk, there is still some risk of a chargeback if a cardholder simply withholds payment, claiming to be a victim of fraud.[27] And, the merchant bears the loss if fraud actually occurs.[28]

Admittedly, even at this level of abstraction the analogy between the circulation of negotiable debt and the processing of on-line card payments is an imperfect one. The clearing of card payments differs from true negotiation in the sense that an on-line merchant can receive value only through a prespecified clearing process and is not free to transfer receipts to any third party she may choose. Also, consumer protection clauses of the Truth in Lending Act keep upstream parties in the clearing process (most notably, the card issuer) from acting as holders in due course of the consumer's debt, that is, enforcing the debt in cases of fraudulent or disputed transactions. Nonetheless, the most characteristic feature of transactions with negotiable instruments—the allocation of credit or fraud risk to the merchant—is shared by on-line transactions using credit cards.

As was the case with the endorsement of circulating bills of exchange, this risk allocation can be defended as a reasonable trade-off between the merchants' need for an on-line payment medium and the credit card issuers' need to contain the risks associated with on-line transactions. According to the least-cost avoider principle, such an allocation makes sense as long as (1) the merchant has some informational advantage over upstream parties in dealing with her customers and (2) the merchant has sufficient wealth at stake to make her endorsement of the customers' debt a meaningful action. These two requirements may be reasonable ones for large on-line retailers but are less likely to be valid for smaller businesses or individuals.[29] And in some instances—for example, in international transactions—even large retailers are reluctant to accept on-line credit card payments.

19. In principle, the Bank of England could have bought the debt of parties other than the government. But the government, as the largest potential debtor, was arguably the best choice for an initial issue of notes.
20. See Bodenhorn (2000). New England was an exception; see Lameroux (1994).
21. At the time, incorporation of a bank required an act of Parliament. The reluctance of Parliament to allow the incorporation of additional banks was no doubt in part motivated by the efficiency of the Bank of England as an engine of government finance.
22. See, for example, Timberlake's (1978) discussion of small-denomination banknotes in the late-nineteenth-century United States.
23. A survey cited in *The Economist* (2001) puts credit cards' share of on-line transactions at 95 percent.
24. Throughout this section the term "on-line merchant" may refer to anyone selling, or wishing to sell, a good or service over the Internet.
25. These parties might include the merchant's bank, a "merchant acquirer" that processes the payment, a credit card company that sets the rules for clearing and settlement of card payments, and the institution that issued the card used to make the payment.
26. In other words, for expositional purposes, assume that the merchant is dealing with a zaibatsu that provides him with supplies, trade credit, and card payment services.
27. This phenomenon is especially prevalent with on-line gambling and adult entertainment services, where it goes by the name of "friendly fraud."
28. Of course, the merchant also has the option of simply not shipping the good until the payment becomes final. For example, an on-line merchant may demand a check payment and wait until the check arrives and clears before shipping merchandise. But this practice may result in a costly and unacceptable delay.
29. Recently a number of firms have begun to offer fraud-detection services for on-line retailers; see, for example, Richmond (2003). These services have no doubt decreased the cost of detecting on-line frauds, but their cost may still be too high for some low-volume on-line merchants.

The existence of would-be on-line purveyors who are unwilling or unable to accept the risks associated with card-based on-line payment has created a demand for alternative payment arrangements. This need has been especially strong for on-line auction sites, where many of the merchants are either households or low-volume retailers. A number of business models have attempted to fill this on-line payment niche.[30] To date, the most successful has been the "on-line currency" arrangement, which offers payment by on-line transfer of a debt claim on a private party. The remainder of this section focuses on the design of the most widely used on-line currency, PayPal, bearing in mind that some of its features might be shared by other on-line currency arrangements.[31]

The most characteristic feature of transactions with negotiable instruments—the allocation of credit or fraud risk to the merchant—is shared by on-line transactions using credit cards.

PayPal works much like an early deposit bank.[32] Deposits are made by transfer of funds to PayPal either by credit card or through electronic funds transfer (debit of the depositor's bank account through the automated clearinghouse [ACH]). Any PayPal account holder can transfer funds to anyone with an e-mail address: both transacting parties are in effect electronically brought to PayPal, and the transfer is made *in banco*. A payer (who has deposited sufficient funds in his account or who has a sufficient line of credit with his credit card company) initiates a payment by visiting the PayPal Web site and typing in the name and address of the payee. The payee receives an e-mail informing him of the payment and has several methods by which he can access the transferred funds. These include (1) circulation, meaning use of the funds received to make additional on-line payments; (2) transfer of funds to the payee's bank account via an ACH credit transfer or by check; or (3) access through an automated teller machine debit card (issued by a bank affiliated with PayPal).

The finality of the on-line transfers is not automatic. First, since deposits to an on-line account are payments made by credit card or through electronic funds transfer (ACH debit), the finality of these transfers falls under the relevant laws and regulations for such transfers.[33] Accordingly, these transfers may be reversed in cases of fraud, or, for credit cards, in cases in which the cardholder authorized the transaction but claims the goods delivered were nonexistent or otherwise defective. The losses from these chargebacks must then be shared between the merchant and PayPal. Unless there is a prior agreement with the merchant (one such agreement is described below), PayPal must attempt to recover the loss from the merchant. Second, according to its IPO filing, PayPal considers transfers on its own books as subject to the Electronic Funds Transfer Act and Regulation E. Under these rules, the liability of account holders is limited in cases of unauthorized transactions, and such limitations can lead to reversals of funds transfers and attempts to recover losses from the payee (merchant).

However, PayPal does provide some assurances of payment finality in cases in which a transaction is covered by its seller-protection policy. Among the requirements are the following:

- The seller has been verified as legitimate.
- Goods are shipped to a verified buyer's address.
- The seller can provide proof that the goods in question have been shipped (intangible goods are therefore not eligible).
- Only one payment has been accepted for the goods in question.
- The goods are shipped to a U.S. buyer at a U.S. address.

If these conditions are met, PayPal assumes risk from unauthorized and false claims of nonshipment of up to $5,000 per year in return for a small fixed fee plus a percentage of each transaction.

Formally, this type of transaction can be thought of as involving the debt swap depicted in Figure 2. The on-line merchant, B, trades the risky obligation of the buyer, C, for the less risky obligation issued by an on-line currency provider, E. The on-line currency provider benefits by charging the merchant a fee on (in other words, by discounting) the transaction but also ends up bearing a good portion of the fraud risk. Hence, for this arrangement to work, the provider must have sufficiently good information on the legitimacy of the buyer.

Do on-line currency providers have access to such information? In the case of PayPal, fraud-loss figures reported in connection with its IPO filing (U.S. Securities and Exchange Commission 2001, 26) indicate that such losses amounted to 0.87 percent of total payment volume in the year 2000 and 0.41 percent of payment volume in the first six months of 2001. These figures compare unfavorably with fraud

rates for traditional payment systems (0.2 percent or less) but are reasonably close to reported figures for on-line credit card fraud (which range from about 0.5 percent to higher than 2 percent). According to press reports (for example, Stone 2001), PayPal's user-verification procedure has been instrumental in containing fraud. This procedure involves depositing small, random amounts of funds in each user's bank account and requiring the user to correctly report the amounts deposited.[34]

Numerous articles in the popular and trade press have documented that on-line currency arrangements have become extremely popular for certain types of on-line payments, most notably for on-line auction sites such as eBay.[35] For such transactions, the appeal of on-line currencies is understandable, particularly in cases where a finality guarantee is provided. In effect, such a system offers a seller in on-line auctions the same "insurance" service that Edward Backwell offered merchants in seventeenth-century London. That is, the idea behind on-line currency provision is to profit from swapping payor obligations for the obligations of the currency provider—in other words, from guaranteeing the transfer of discounted buyers' claims to sellers.

As was the case with early banknotes, the revenue stream from providing an on-line currency is principally derived from discounting—from transactions fees—rather than from collecting interest on funds circulating as on-line balances. For example, in the second quarter of 2001, PayPal had revenues of $19.9 million, with $18.6 million, or 94 percent, derived from transactions fees. In other words, there is little evidence so far that people are using on-line currencies for any purpose other than one-time, on-line purchases.

Caveats

The discussion above has laid out one possible explanation for the popularity of on-line currencies—their ability to provide finality in on-line transactions. Kuttner and McAndrews (2001) and Schreft (2002) lay out some alternative explanations for the popularity of on-line currency payments. First among these is convenience. Since payments can be made by e-mail, there is no need for a household or very low-volume merchant to set up a merchant account to receive on-line currency payments. For higher-volume merchants, however, the opposite consideration may hold true; that is, the merchant may prefer to consolidate all on-line payments through a single payment processor, usually a firm that processes credit card payments.

Price may be another significant factor behind the use of on-line currencies. To date, on-line currency payments have typically been free (or close to it) for individuals and merchants with low transaction volumes. For these people, accepting on-line currency payments is simply cheaper than accepting credit card payments. For slightly higher transaction volumes, however, transaction fees are charged, and published merchant per-transaction fees for PayPal are roughly comparable to those posted for on-line credit card payments. This detail suggests that other factors determine merchants' decisions about which types of payments to accept. These factors could include the effort and expense of maintaining a merchant account for receiving credit card payments. But for many low-volume merchants, finality considerations may be the most important factor: For small operations, just a few significant losses due to on-line fraud could easily negate the benefits of a lower transaction fee.[36]

While the finality rationale for the provision of on-line currencies closely mimics that of the earliest banknotes, some marked differences in the function and implementation of the two sets of arrangements are also apparent. Chief among these differences is that, to date, on-line currencies have been

30. Kuttner and McAndrews (2001) and Schreft (2002) survey various modes of on-line payment.
31. PayPal, which was launched in 1999, was recently (October 2002) acquired by the on-line auction company eBay. Basic information on PayPal is available from the prospectus filed for its initial public offering (IPO) (U.S. Securities and Exchange Commission 2001) and from its Web site, www.paypal.com.
32. As of this writing, however, PayPal does not offer traditional checking accounts and is not legally recognized as a bank. PayPal is considered a "money transmitter" under the laws of many states.
33. These laws and regulations include but are not necessarily limited to the Truth in Lending Act/Federal Reserve Regulation Z for credit card payments and the Electronic Funds Transfer Act/Federal Reserve Regulation E for funds transfers.
34. PayPal has applied for a patent on this procedure, but, in the meantime, other providers of on-line payment services may be using it.
35. See, for example, Sapsford and Beckett (2001) or Slatalla (2001). Wingfield and Sapsford (2002) report that 70 percent of electronic transactions over eBay are PayPal transactions.
36. This view is consistent with press reports concerning the hazards confronting small businesses or households selling on-line. See, for example, Richmond (2003).

used primarily for relatively small-value transactions.[37] The cost of managing the risks associated with these transactions is high relative to the value of payments transferred, and whether these costs can be kept within reasonable limits over the longer haul is an open question. On-line currency providers have moved aggressively to contain fraud risk—by implementing the verification procedures described above, by placing limits on transaction balances, employing pattern-recognition programs to detect fraud, and freezing accounts suspected of fraudulent activity. Such aggressive strategies can easily backfire, however, by undermining confidence in the very service that the currency providers are attempting to sell—that is, payment finality.

The need to conduct transactions with strangers over the Internet has created a demand for new payment technologies, as did the need to conduct transactions over distance with strangers three hundred years ago.

A related issue is the degree of finality that is appropriate for small-value on-line transactions. Payment finality insulates the seller from the risk of fraud on the buyer side but also creates an incentive for the seller to provide substandard merchandise or simply no merchandise at all. Compared to the wholesale payment environment that spawned the first banknotes (where one might reasonably expect all parties to have been well informed about the risks involved and reasonably capable of bearing these risks), in an on-line environment, providing buyers with some degree of recourse or insurance against fraudulent sales may be desirable for transactions involving consumers. At least one on-line currency provider (PayPal) insures buyers to a limited extent against fraud, but such insurance is again likely to raise the cost of providing the on-line currency.

Another fundamental distinction between the on-line currencies and the early banknotes is that, thus far, on-line currencies have been provided by stand-alone enterprises only tangentially connected to mainstream banking and payment industries. No bank, card association, or other payment-card issuer has offered an on-line currency up to now. On-line currency providers have thus been unable to take advantage of potential economies of scope in managing information about their customers.[38]

On-Line Currencies: Prospects

What is the future of on-line currencies? Certainly the on-line currency business model has achieved a measure of success in its role as payment provider for on-line person-to-person and consumer-to-small-business transactions. But even within this niche, on-line currency providers face competition from other payment technologies. An important question is whether on-line currencies will be supplanted by these other technologies in person-to-person applications.

One source of competition is bank-affiliated on-line payment systems such as c2it (operated by Citibank). These bank-sponsored systems essentially function as facilitators for traditional payment systems, allowing on-line buyers to send funds either by credit card or ACH debit to on-line sellers. While the bank-sponsored systems do not offer finality guarantees (beyond the guarantees of the underlying payment mechanism), their affiliation with banks confers some potentially important advantages in the provision of payment services. These advantages include automatic access to existing payment and settlement systems, extensive information on the banks' own customers, a widespread physical infrastructure, and a wealth of human capital in managing risks associated with small-value payments. And since banks are regulated institutions, consumers may feel that payment services associated with banks are safer and more likely to be cooperative in resolving disputes even if the full extent of the safety net of bank regulation may not extend to banks' associated on-line payment systems.

Credit card companies are also trying to lower fraud risk in on-line payments, for example, by making use of software that generates credit card numbers that can be used for only a single on-line purchase. Credit card holders can use one of these single-use card numbers without having to reveal the number of their physical credit card, thereby lessening the chance that their card number will be put to fraudulent use. More recently, credit card companies have introduced on-line authentication systems (see, for example, Punch 2002) in an attempt to control on-line fraud. Merchants who use these systems require would-be on-line purchasers to first obtain a unique password from their credit card company. The purchaser must then enter the password before using his credit card to make the on-line purchase. The use of these systems and other technological improvements in credit card payments may eventually result in more widespread acceptance of credit cards by low-volume on-line retailers and hence less demand for payment by on-line currency.

TABLE

Parallels in the Development of Banknotes and On-Line Currencies

	Circa 1700	Circa 2000
Problem: Arranging transactions between strangers	Transactions must be made over distance	Transactions must be made over the Internet
Solution: Third-party instruments (merchant transfers the debt of a customer to pay off existing debt)	Bills of exchange (merchant transfers the debt of a customer by "drawing a bill" on him)	Credit cards (merchant transfers a consumer's credit card payment)
Problem: Adverse selection ("lemons") problem of low-quality payments driving out high-quality payments	Bills are forged or drawn on poor credits	Credit card fraud
Solution: Merchant signals quality of debt by accepting contingent liability	A merchant must endorse a bill before transferring it	A merchant must accept chargeback liability for on-line credit card payments
Problem: Contingent liability excludes some merchants	The value of an endorsement depends on reputation and collateral	The cost of chargeback liability is too high for some low-volume merchants
Solution: Third-party supply of liability	A London goldsmith discounts a bill of exchange and in return issues a bearer note or banknote	An on-line currency issuer accepts a credit card payment and creates a claim that can be transferred on-line
Opportunity: Expand innovation beyond niche market	Bank of England buys a large issue of government debt and issues many banknotes; these come into general circulation	?

A more fundamental question, and one even more difficult to answer, is whether on-line currencies will break out of the on-line person-to-person niche and become a widely accepted substitute for more traditional forms of payment such as checks and currency. The volume of payments made through on-line payment providers remains relatively small, probably below 500,000 per day according to Kuttner and McAndrews (2001), as compared to more than 200 million payments made daily through traditional payment instruments. Just as users of early banknotes did, on-line currency issuers face a chicken-and-egg problem: liquidity (and hence profitability from circulation) is linked to the scale of the currency issue, and, conversely, the scale is linked to liquidity. No doubt the demand for on-line currencies has been limited by the inconvenience of converting traditional forms of inside money (bank deposits) to an on-line currency and vice versa. Demand for on-line currencies could be enhanced if people could receive an on-line currency payment with the expectation that the currency could be passed on without having to convert it into traditional bank money. To date, no on-line currency provider has been successful in creating such an expectation.

Conclusion

This article has argued that there are certain parallels between the current on-line payment environment, which led to the development of on-line currencies, and the physical payment environment of roughly three hundred years ago, which led to the debut of banknotes. These parallels are summarized in the table.

The first parallel is in the emergence of a demand for new payment technologies. The need to conduct transactions with strangers over the Internet has created such a demand, as did the need to conduct transactions over distance with strangers three hundred years ago. In response to this demand,

37. For the second quarter of 2001, PayPal reported that 85 percent of its transactions were below $1,000 and that the average transaction value was about $50. See U.S. Securities and Exchange Commission (2001, 1).
38. However, the on-line verification system used by PayPal and some other on-line payment providers (discussed above) does in effect free ride on information gathered by banks concerning their customers.

payment providers in both cases established new payment technologies—on-line credit card payments now and negotiable bills then.

The second parallel is in the problem of adverse selection in these new types of payments, particularly over the risk of fraud on the buyer side of the transactions. The solution, in both the current and historical cases, has been to provide limited finality and concentrate fraud risk on sellers (through credit card chargebacks now and endorsement then) who accept payments that use the new types of payment technologies.

The third parallel derives from the fact that this last risk allocation is not always the most desirable one for all transactions. Households and low-volume merchants in particular may be unable or unwilling to bear the risk of buyer-side fraud. The solution has been to create a new type of payment technology that allocates much of the buyer-side fraud risk to an outside party—on-line currency providers now, goldsmith issuers of banknotes then.

Despite these evident parallels, we would stop short of calling on-line currencies "virtual banknotes," at least for the time being. This hesitance exists because the final step in on-line currencies' "monetization"—widespread acceptance as a circulating medium of exchange—has yet to occur. It remains to be seen whether an on-line currency issuer will overcome the financial and technical, not to mention legal and regulatory, hurdles associated with scaling up its on-line currency into a viable competitor to traditional payment media. Whether such a feat—comparable to the Bank of England's initial banknote issue—is possible in today's world is, at best, debatable. But if monetary history is any guide, the resulting payoff would be large.

REFERENCES

Bank for International Settlements. Committee on Payment and Settlement Systems. 2001. *Statistics on payment systems in the Group of Ten countries.* Basel: BIS.

Berger, Allen N., Diana Hancock, and Jeffrey C. Marquardt. 1996. A framework for analyzing efficiency, risks, costs, and innovations in the payments system. *Journal of Money, Credit, and Banking* 28, no. 4, part 2:696–732.

Bodenhorn, Howard. 2000. *A history of banking in antebellum America: Financial markets and economic development in an era of nation-building.* Cambridge: Cambridge University Press.

Boeschoten, Willem. 1992. Currency use and payment patterns. *Financial and monetary policy studies*. Vol. 23. Norwell, Mass.: Kluwer Academic Publishers.

Checkland, Sydney G. 1975. *Scottish banking: A history, 1695–1973*. London: Collins.

Committee on the Federal Reserve in the Payments Mechanism. 1998. *The Federal Reserve in the payments mechanism.* Washington, D.C.: Federal Reserve Board of Governors. <www.federalreserve.gov/boarddocs/press/general/1998/19980105/19980105.pdf> (February 1, 2003).

Dehing, Pit, and Marjolein 't Hart. 1997. Linking the fortunes: Currency and banking, 1550–1800. In *A financial history of the Netherlands*, edited by Marjolein 't Hart, Jost Jonker, and Jan Luiten van Zanden. New York: Cambridge University Press.

De Roover, Raymond. 1948. *Money, banking, and credit in mediaeval Bruges*. Cambridge, Mass.: Mediaeval Academy of America.

DeRosa, Luigi. 2001. The beginnings of paper money circulation and Neapolitan public banks (1540–1650). *Journal of European Economic History* 30 (Winter): 497–532.

The Economist. 2001. E-cash 2.0, February 17.

Horsefield, J. Keith. 1983. *British monetary experiments, 1650–1710*. New York: Garland Publishing.

Humphrey, David B. 2002. U.S. cash and card payments over 25 years. Florida State University. Manuscript. <www.phil.frb.org/econ/conf/innovations/Humphrey.pdf> (February 1, 2003).

Kahn, Charles M., and William Roberds. 2001. Transferability, finality and debt settlement. Federal Reserve Bank of Atlanta Working Paper 2001-18a, October. <www.frbatlanta.org/invoke.cfm?objectid=16E346B9-C490-11D5-A37F0008C7720D25&method=display> (February 1, 2003).

Kuttner, Kenneth N., and James J. McAndrews. 2001. Personal on-line payments. Federal Reserve Bank of New York *Economic Policy Review* 7 (December): 35–50.

Lameroux, Naomi. 1994. *Insider lending*. New York: Cambridge University Press.

Lee, W.A. 2003. Fraud keeps on increasing as thieves romp on the Web. *American Banker*, January 31.

Mann, Ronald J. 1999. *Payment systems and other financial transactions.* New York: Aspen Law and Business.

Mengle, David L. 1990. Legal and regulatory reform in electronic payments: An evaluation of payment finality rules. In *The U.S. payment system—efficiency, risk, and the role of the Federal Reserve: Proceedings of a symposium on the U.S. payment system sponsored by the Federal Reserve Bank of Richmond*, edited by David B. Humphrey. Boston: Kluwer Academic Publishers.

Mueller, Reinhold C. 1997. *The Venetian money market: Banks, panics, and the public debt, 1200–1500.* Baltimore: Johns Hopkins University Press.

Munro, John. 2000. English "backwardness" and financial innovations in commerce with the Low Countries, 14th to 16th centuries. In *International trade in the Low Countries (14th–16th Centuries): Merchants, organisation, infrastructure, studies in urban, social, economic, and political history of the medieval and early modern Low Countries*. Vol. 10. Edited by Peter Stabel, Bruno Blondé, and Anke Greve (Marc Boone, general editor). Leuven-Apeldoorn, The Netherlands: Garant.

Neal, Larry. 1991. *The rise of financial capitalism.* New York: Cambridge University Press.

Porter, Richard D., and Ruth A. Judson. 1996. The location of U.S. currency: How much is abroad? *Federal Reserve Bulletin* 82 (October): 883–903.

Pressnell, Leslie. 1956. *Country banking in the Industrial Revolution*. Oxford: Clarendon Press.

Punch, Linda. 2002. Authentication's tentative gains. *Credit Card Management*, April 25.

Quinn, Stephen F. 1997. Goldsmith-banking: Mutual acceptance and interbanker clearing in Restoration London. *Explorations in Economic History* 34 (October): 411–32.

Richards, Richard David. 1929. *The early history of banking in England*. New York: A.M. Kelley.

Richmond, Riva. 2003. Scammed! Web merchants use new tools to keep buyers from ripping them off. *Wall Street Journal*, January 27.

Royal Bank of Scotland. 2003. *The ledgers of Edward Backwell—banker of London.* CD-ROM. Edinburgh: Privately published by the Royal Bank of Scotland.

Sapsford, Jathon, and Paul Beckett. 2001. Credit-card firms still need a strong hand in Web. *Wall Street Journal*, April 2.

Schreft, Stacey. 2002. Clicking with dollars: How consumers can pay for purchases from e-tailers. Federal Reserve Bank of Kansas City *Economic Review* 87 (First Quarter): 37–64.

Slatalla, Michelle. 2001. Easy payments put hole in the pocketbook. *New York Times*, June 29.

Smith, Adam. [1776] 1994. *An inquiry into the nature and causes of the wealth of nations.* Reprint, New York: The Modern Library.

Stone, Brad. 2001. Busting the online bandits. *Newsweek*, July 17.

Timberlake, Richard H. 1978. *Origins of central banking in the United States.* Cambridge, Mass.: Harvard University Press.

U.S. General Accounting Office. 1997. *Payments, clearance, and settlement: A guide to the systems, risk, and issues*. Washington, D.C., June. <www.gao.gov> (February 1, 2003).

Usher, Abbot P. 1943. *The early history of deposit banking in the Mediterranean.* Vol. 1 of *Structure and functions of the early credit system and banking in Catalonia, 1240–1723.* Cambridge, Mass.: Harvard University Press.

U.S. Securities and Exchange Commission. 2001. Form S-1 Registration Statement under the Securities Act of 1933: PayPal, Inc. Filed September 28.

Van der Wee, Herman. 1997. The influence of banking on the rise of capitalism in northwest Europe, fourteenth to nineteenth century. In *Banking, trade, and industry*, edited by Alice Teichova, Ginette Kurgan-van Hentenryk, and Dieter Ziegler. New York: Cambridge University Press.

White, Lawrence H. 1984. *Free banking in Britain*. New York: Cambridge University Press.

Wingfield, Nick, and Jathon Sapsford. 2002. EBay to buy PayPal for $1.4 billion. *Wall Street Journal*, July 9.

Winn, Jane Kaufman. 1998. Couriers without luggage: Negotiable instruments and digital signatures. *South Carolina Law Review* 49. <www.law.washington.edu/Faculty/Winn/Publications/Couriers%20without%20Luggage.htm> (February 1, 2003).

Forecast Evaluation with Cross-Sectional Data: The Blue Chip Surveys

ANDY BAUER, ROBERT A. EISENBEIS, DANIEL F. WAGGONER, AND TAO ZHA

All of the authors work in the Atlanta Fed's research department. Bauer is an analyst, Eisenbeis is senior vice president and director of research, Waggoner is an economist and assistant policy adviser, and Zha is an assistant vice president and policy adviser.

The authors thank Bevin Janci for her valuable research assistance.

Evaluating the accuracy of economic forecasts is critical if they are to be used in decision making. When a single variable is being forecast by a model with several independent variables, accuracy is typically evaluated using mean square error, mean absolute error, or some similar criterion. When a set of variables is being projected in a simultaneous equation setting, researchers still typically assess forecast accuracy for each dependent variable in the system separately. Though this practice is acceptable as a first pass, it ignores three important aspects of the accuracy assessment process. First, using univariate comparisons to forecast a joint system fails to consider possible correlations in the forecast errors, which might bias the assessment. Second, univariate approaches may not be able to rank forecasters uniquely in terms of their overall performance because one forecaster may perform better on one variable while others may perform better on other variables. This consideration is important because these forecasters are projecting a set of economic variables, which should be internally consistent. Being off on several key dimensions but right on one variable provides some indications about the overall quality of the forecast. Finally, while currently employed statistical comparisons reveal how well models or forecasters may perform on average, they do not help to evaluate and compare particular point forecasts at given times.

Using the methodology developed in Eisenbeis, Waggoner, and Zha (2002), which addresses each of the problems mentioned previously, this article explores and compares the economic forecasts in the Blue Chip Economic Indicators Survey. These data are particularly well suited for the problem at hand. The survey has been published monthly since 1977 and contains forecasts of many macroeconomic variables over a relatively long time span. Although variables have been added or dropped, a substantial number have been present since the survey's inception. The forecasters are a mix of economists from major investment banks, corporations, consulting firms, and academic institutions. On average, the survey contains fifty forecasts each month, and many of the forecasters have participated in the survey for several years. The survey thus provides a useful set of forecasts to explore the methodologies and to investigate several aspects of forecast performance over time.

The article also examines whether several key assumptions underlying the measures advocated in Eisenbeis, Waggoner, and Zha (2002) hold; the results show that these assumptions are satisfied for the Blue Chip data set, at least for longer horizon forecasts. The analysis shows that the Blue Chip Consensus Forecast, which is the average of the individual forecasts, performs better than any individual forecaster although several forecasters performed almost as well as the consensus. This

finding indicates that averaging the forecasts across many forecasters removes some of the noise in each individual forecast. This finding also has implications for combining forecasts from different econometric models, a practice that has been extensively explored in the literature (Bates and Granger 1969; Newbold and Granger 1974; Clemen 1989; de Menezes, Bunn, and Taylor 2000).

The discussion first outlines the methodology used in Eisenbeis, Waggoner, and Zha (2002) and details the Blue Chip data and the benchmark data used to evaluate the forecasts. The article then describes the empirical results and provides some conclusions.

The rank and the score of a forecast are similar in the sense that over time the two measures will be uniformly distributed over some interval.

Methodology

There are many different ways to assess the accuracy of forecasts. Ultimately, determining which forecast is best depends on the use to which it will be put. If accuracy in forecasting output is more important than accuracy in forecasting inflation, then one will want to use forecasts that deliver accurate measures of output relative to inflation. The purpose of this article is to evaluate and compare the general accuracy of a set of multivariate forecasts over time. The methodology in its basic form penalizes errors on easy-to-forecast dimensions more than errors on hard-to-forecast dimensions and considers correlations among the forecast errors.[1] Following Eisenbeis, Waggoner, and Zha (2002), this study uses a composite score based on the standard theory of probability and statistics. This score can be used to compare forecasts even if the number of variables being forecast, or their definitions, changes over time. Finally, the method has the advantage of reducing forecast performance assessment to a single number with an easy interpretation.

In one dimension, the squared error is a standard choice to evaluate and compare forecasts. For example, if y is gross domestic product (GDP) growth and $\hat{y}$ is a forecast of y, then $(y - \hat{y})$ is the forecast error and $(y - \hat{y})^2$ is the squared error. If the forecast error is normal with mean zero and variance σ^2, then the normalized squared error,

$$(1)\quad (y - \hat{y})^2/\sigma^2,$$

has a chi-square distribution with one degree of freedom.[2] With the aid of a chi-square table, one could look up the probability of observing a normalized squared error even larger than the given one. This theoretical probability, converted to a percentage, is the score of a forecast as defined in Eisenbeis, Waggoner, and Zha (2002). Under the assumption of normality, over time the forecast scores would vary uniformly between 0 and 100.[3] This assumption does not mean that the scores of each forecaster would vary uniformly between 0 and 100. A superior forecaster might have scores concentrated in the upper end of this range while an inferior forecaster's scores might lie mostly in the lower end.

To evaluate multivariate forecasts, one simply uses the multivariate generalization of the univariate normal distribution. (See the box on page 20 for a discussion of the multivariate normal distribution.) Suppose that $\mathbf{y}$ is a vector of economic variables to be forecast. For the sake of illustration, suppose that $\mathbf{y}$ consists of only two variables: GDP and the consumer price index (CPI). If $\hat{\mathbf{y}}$ is a forecast vector of the two variables, then the forecast error is the vector of the difference between forecast and realized GDP and forecast and realized CPI, denoted by $(\mathbf{y} - \hat{\mathbf{y}})$. If the forecast error has a multivariate normal distribution with mean 0 and variance $\mathbf{\Omega}$, then the analog of (1) is

$$(2)\quad (\mathbf{y} - \hat{\mathbf{y}})'\mathbf{\Omega}^{-1}(\mathbf{y} - \hat{\mathbf{y}}),$$

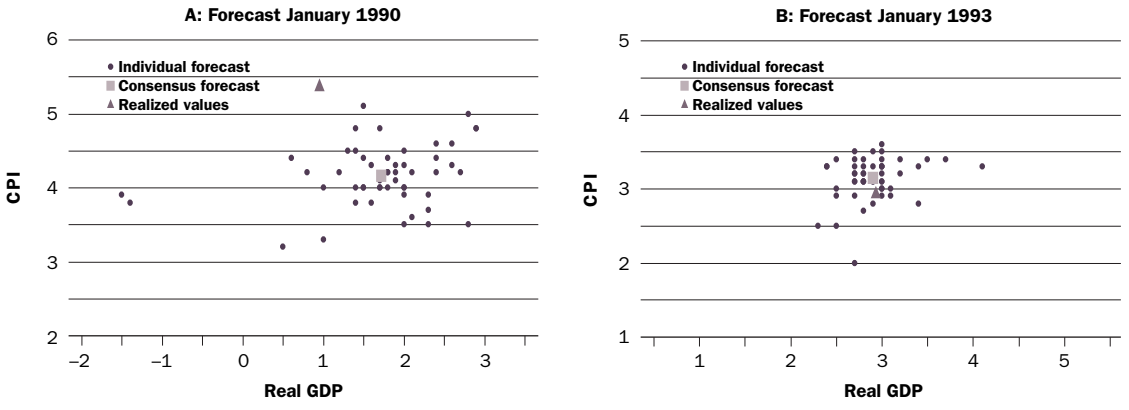
which has a chi-square distribution with degrees of freedom equal to the number of variables.[4] The score of a forecast is defined as the probability, in percentage terms, of observing a normalized squared error even larger than the given one.

The normalized squared error given by equation (1) or (2) is a special case of the more general notion of a loss function. A loss function is simply a mapping of the forecast errors to non-negative numbers and is interpreted as the loss, economic or otherwise, resulting from making a particular error. If the distribution of the loss function applied to the forecast errors were known, then the score of a forecast could be defined as the probability of observing a loss even greater than the loss associated with the given forecast error. This approach is used in Eisenbeis, Waggoner, and Zha (2002).

Loss functions can also be used to rank forecasts. The rank and the score of a forecast are similar in the sense that over time the two measures will be uniformly distributed over some interval; if a forecaster

FIGURE 1

Forecasts of Real GDP and CPI



has superior (or inferior) skill, then the measures for that forecaster will be skewed toward one end of the interval. The difference between these measures is that the rank is always relative to the other forecasters in the group while the score is in absolute terms. If the realized value of the forecast variables is far from the average forecast, then most of the scores will be low while the ranks will always be distributed between 1 and the number of forecasters. Both measures are useful, and both will be reported in this article.

In some contexts there is an obvious candidate for the loss function, but in general there is often no canonical choice. The loss functions given by equations (1) and (2) are called quadratic loss functions and have often been used in the forecasting literature. In univariate models, the choice of σ will have no effect on the forecasts' ranking. However, in multivariate models different choices of Ω will induce different rankings. The matrix Ω determines, among other things, the relative importance of the forecast errors of the individual variables. Assigning different weights to these forecast errors could produce different ranks. This analysis uses assumptions about the distributions of the forecast errors to inform the choice of Ω. Errors in forecasting easy-to-forecast variables are penalized more than errors in forecasting hard-to-forecast variables.

The forecast error variance can be divided into two segments—error variance attributed to unpredictable events and error variance caused by using imperfect forecasting models, also known as model uncertainty. Even if a forecaster had access to all available information at a given time and had perfect foresight in combining this information to make a forecast, the forecast usually would not be equal to the observed values. Unpredictable events occurring after the forecast has been made ensure that no forecast of economic variables can always be exact. The variance of this hypothetical best forecast relative to the realized value is denoted by Ω^H.[5] The consensus forecast can be used to approximate the hypothetical best forecast.[6]

In practice, a forecaster does not have access to all available information or perfect foresight in using information to make forecasts. Thus, an actual forecast will vary from this hypothetical best forecast. This variance is denoted by Ω^F. Figure 1 compares the joint forecasts of GDP and the CPI for two arbitrarily chosen periods.

1. The methodology could easily be generalized to consider the costs of different types of errors.
2. Information about and tables for the chi-square distribution can be found in any elementary statistics text. It is important to note that this definition depends heavily on the assumption of normality of the forecast error. In practice, the forecast error is not exactly normal but is close enough so that this is not an extreme assumption.
3. If x is a normalized squared error and s is its associated score, then the probability of observing a score less than s is equal to the probability of observing a normalized squared error greater than x, which, from the definition of the score, is $s/100$. Thus, the probability of observing a score in any subinterval of (0, 100) is proportional to the length of the subinterval. This proportionality is the defining feature of the uniform distribution.
4. Here, $(\mathbf{y} - \mathbf{y})'$ is a row vector, Ω^{-1} is the matrix inverse of Ω, and $(\mathbf{y} - \hat{\mathbf{y}})$ is a column vector. The product $(\mathbf{y} - \hat{\mathbf{y}})'\Omega^{-1}(\mathbf{y} - \hat{\mathbf{y}})$ is matrix multiplication and results in a single number.
5. In this case, the hypothetical best forecast is also known as the conditional mean of the variables being forecast. The conditional mean minimizes the expected score.
6. Using the language of Blue Chip, the consensus forecast is considered to be the mean, or average, forecast.

BOX

Multivariate Normal and Chi-Square Distributions

A univariate normal distribution is characterized by two numbers, the mean and the variance. The mean centers the distribution, and the variance determines the dispersion. A multivariate normal distribution is also characterized by its mean and variance, but the mean is a vector and the variance is a matrix. The figure shows a sample of 200 points from a two-dimensional normal distribution. The mean of this distribution is (1, 2), and the variance is

$$\begin{pmatrix} 0.81 & 0.27 \\ 0.27 & 0.36 \end{pmatrix}.$$

Each coordinate of a multivariate normal distribution will have a univariate normal distribution. In this case, the mean of the first coordinate is 1 with a variance of 0.81, and the mean of the second coordinate is 2 with a variance of 0.36. The covariance of the two coordinates, which is the correlation times the square root of the variances, is 0.27. Thus, the correlation is $0.27/\sqrt{0.81 \times 0.36} = 0.5$. In general, the elements of a variance matrix are the variance and covariance among the individual coordinates.

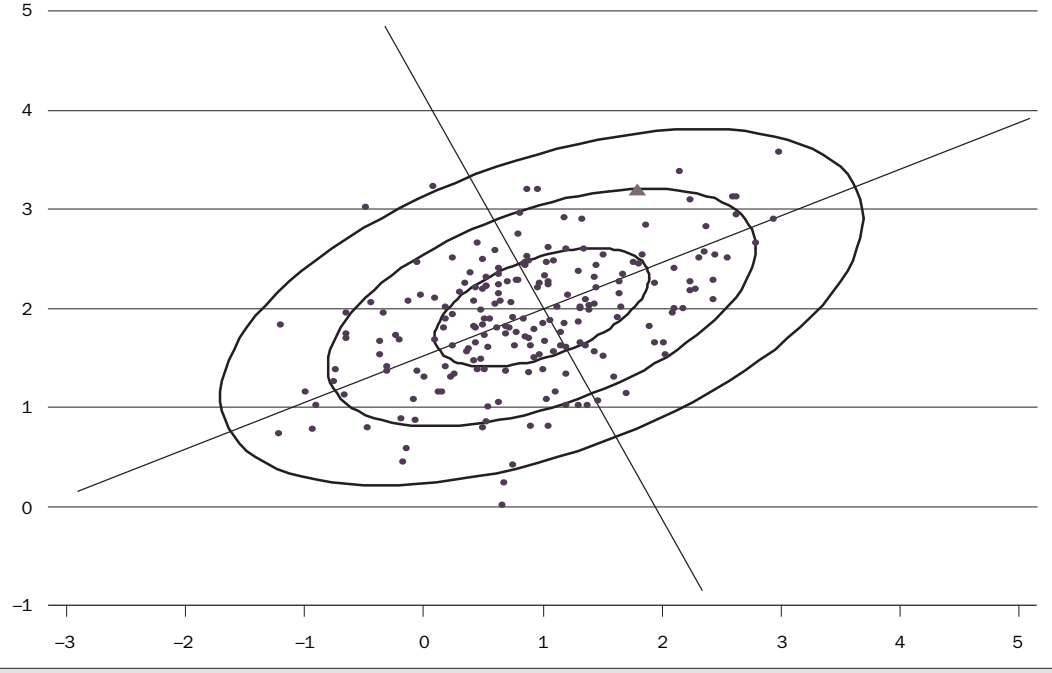
The triangular point on the middle ellipse has coordinates (1.8, 3.2), and the matrix product

$$(1.8 - 1 \;\; 3.2 - 2)\begin{pmatrix} 1.65 & -1.23 \\ -1.23 & 3.70 \end{pmatrix}\begin{pmatrix} 1.8 - 1 \\ 3.2 - 2 \end{pmatrix}$$

is approximately 4. The row and column vectors are the difference between the triangular point and the mean while the matrix is the inverse of the variance given above. The middle ellipse has the property that the above product will always be 4 for any point along the ellipse. For points inside this ellipse, the product will be less than 4, and for points outside the ellipse the product will be greater than 4. On the outer ellipse the product will be 9, and on the inner ellipse the product will be 1. By applying the above product to any data point, a two-dimensional normal distribution is transformed into a one-dimensional chi-square distribution with two degrees of freedom. In general, the degrees of freedom will be equal to the number of variables. Indeed, the chi-square distribution is defined in this way from a multivariate normal distribution. Using a chi-square table (or the Microsoft Excel function CHIDIST), one can find the probability that a data point will lie outside a given ellipse or, equivalently, the probability that the product computed from the data point will be greater than some given value. This method defines and should be used to interpret the score of a forecast.

FIGURE

Sample from a Two-Dimensional Normal Distribution



In both panels A and B there is considerable variation between the individual forecasts and the consensus. This variation is captured by Ω^F. The difference between the consensus forecast and the realized values is fairly large in panel A but is smaller in panel B. This variation is captured by Ω^H. The total variation over time will be the sum of these two.

Symbolically, if $\mathbf{y}$ denotes the vector of economic variables being forecast, $\hat{\mathbf{y}}$ is an individual forecast, and $\overline{\mathbf{y}}$ is the hypothetical best or consensus forecast, then the forecast error can be partitioned as

$$\mathbf{y} - \hat{\mathbf{y}} = (\mathbf{y} - \overline{\mathbf{y}}) + (\overline{\mathbf{y}} - \hat{\mathbf{y}}).$$

The first term is the error in the hypothetical best or consensus forecast, and the second term is the additional error due to the difference between the actual forecast and the best forecast. If these two errors are independent, then the variance of the forecast error will simply be the sum of the two corresponding variances, Ω^H and Ω^F. Thus,

$$\Omega = \Omega^H + \Omega^F.$$

Given the large cross section of data in the Blue Chip Survey, Ω^F can easily be estimated as the variance matrix of the forecasts under the assumption that the Blue Chip Consensus Forecast is an acceptable proxy for the hypothetical best forecast (see recent work by Ottaviani and Sorensen 2003). The matrix Ω^H can be estimated as the variance matrix of the realized forecast errors of the consensus forecast. About fifty forecasts are in each Blue Chip Survey, enough to estimate Ω^F. However, only one Blue Chip Consensus Forecast exists for each period, so a relatively long series of forecasts is needed to estimate Ω^H. There must be at least as many consensus forecasts as there are variables being forecast, and ideally there should be more than three to four times as many consensus forecasts as variables. In Eisenbeis, Waggoner, and Zha (2002), a particular forecasting model was estimated, and Ω^H was taken to be the theoretical variance matrix from this model. This estimation was necessary because the variables used in the *Wall Street Journal* forecasts frequently changed over time, and so there was not a long enough time series of mean forecasts. Both methods for obtaining Ω^H are considered here to compare the sensitivity of the proposed performance measures to differences in estimates of Ω^H.

The Data

The Blue Chip Economic Indicators Survey has been published monthly for more than twenty-five years. The survey includes the annual average of, or change in, fifteen macroeconomic variables. Forecasts of five variables are considered here: real GDP, the CPI, the unemployment rate, three-month Treasury bill rates, and ten-year Treasury note yields. With the exception of the ten-year Treasury note, these variables have been included in all of the surveys. Prior to 1996, a corporate bond yield was forecast instead of the Treasury note. Though differences exist between Treasury and corporate yields, these series are joined for illustrative purposes in this study. Approximately fifty firms participate in each survey. Though the number of firms has remained roughly constant, the identities of the firms have changed over time. In some instances, firms merged or ceased to appear in the survey for various reasons. The analysis tracks merged firms and combined forecasts to create long time series when possible. The participation dates of each firm and mergers are noted in Table 1.

The forecast error variance can be divided into two segments—error variance attributed to unpredictable events and error variance caused by using imperfect forecasting models, also known as model uncertainty.

Understanding the dating conventions of the forecasts is important for understanding the tables, figures, and discussions in this article. Near the beginning of every month each forecaster submits two forecasts: one for the current calendar year and another for the next calendar year. For instance, in January 2000, forecasters submit a forecast for 2000 (current year) and for 2001 (next year). Forecasts are dated by the year and month in which they are made and identified as either current or next year. Next-year forecasts made in January are long-term forecasts. These forecasts will not be completely realized for twenty-four months. On the other hand, current-year forecasts made in December are short-term forecasts that will be realized in one month. For each year being forecast, twenty-four forecasts with horizons varying from one to twenty-four months are made. This study includes current-year forecasts from January 1986 through December 2001 and next-year forecasts from January 1986 through December 2000.

To determine the accuracy of the forecasts, benchmark or realized values of the variables must

TABLE 1

Average Scores

	Years in Survey[1]	Average Score		Current-Year Average Score		Next-Year Average Score	
BC Consensus	86–01	69.3***	(21.8)	74.7***	(23.2)	63.5**	(18.7)
Security Pacific National Bank	86–92	68.8**	(24.4)	73.9**	(27.2)	63.6	(19.8)
NationsBank	93–98	67.7*	(23.0)	67.7*	(27.7)	67.7*	(17.3)
Mortgage Banker Assn. of America	86–01	67.1***	(25.9)	73.2***	(26.4)	60.7*	(23.8)
Macroeconomic Advisors[2]	86–01	66.6**	(25.9)	74.2***	(25.4)	58.4	(24.0)
U.S. Trust Company	86–01	63.6**	(26.0)	68.3***	(25.7)	56.1	(24.8)
CoreStates Financial Corporation	88–98	63.0*	(24.4)	61.4*	(28.2)	64.7**	(19.7)
Pennzoil Company	86–89, 92–93	62.9	(26.4)	68.0*	(28.0)	57.7	(23.9)
Northern Trust Company	86–01	62.7**	(26.6)	66.0**	(26.9)	58.9	(25.7)
Bank of America	87–01	62.7**	(26.2)	64.7**	(28.5)	60.6*	(23.5)
Equitable Life Assurance	86–91	62.5	(26.3)	69.6**	(25.0)	52.4	(25.0)
Peter L. Bernstein, Inc.	86–89	62.2	(28.8)	64.8	(28.5)	59.6	(29.3)
Moody's Investors Service	98–01	61.7	(26.7)	66.5	(31.2)	55.0	(17.2)
Wayne Hummer Investments, LLC	86–01	60.6*	(25.7)	62.5**	(27.5)	58.6	(23.6)
Merrill Lynch	86–01	60.1*	(26.3)	64.2**	(27.2)	55.5	(24.5)
Dean Witter Reynolds & Company	86–91	60.0	(30.0)	63.0	(30.7)	56.1	(29.1)
PNC Financial Corporation	88–98	59.3	(24.3)	59.2	(27.6)	59.3	(20.3)
Fleet Financial Group[3]	91–99	58.8	(24.1)	62.7*	(28.5)	54.9	(18.0)
Metropolitan Life Insurance Company	86–96	58.7	(26.0)	59.7	(31.8)	57.7	(18.6)
Wells Capital Management[4]	91–01	58.2	(27.2)	62.8*	(29.0)	53.1	(24.3)
Georgia State University	86–01	58.2	(26.3)	59.2	(28.1)	57.2	(24.3)
National Association of Home Builders	90–01	58.1	(24.3)	62.8*	(25.7)	53.0	(21.5)
Chicago Capital, Inc.	96–00	57.6	(32.5)	63.3	(31.4)	51.7	(32.8)
University of Michigan M.Q.E.M.	86–96	56.9	(28.7)	67.5**	(28.1)	46.3	(25.3)
National City Corporation[5]	86–01	56.8	(24.0)	59.4*	(25.6)	54.0	(21.9)
Evans Group	86–01	56.5	(28.6)	63.4**	(28.7)	49.1	(26.7)
Eggert Economic Enterprises, Inc.	86–01	56.5	(24.3)	55.6	(27.1)	57.5	(21.1)
DaimlerChrysler AG[6]	86–01	56.3	(28.1)	62.8**	(28.1)	49.4	(26.3)
Chase Manhattan Bank	88–00	56.2	(28.7)	61.3*	(29.1)	50.3	(27.2)
La Salle National Bank	86–91, 97–01	56.1	(28.0)	62.0*	(30.1)	49.4	(24.0)
Dun & Bradstreet	89–99	55.9	(28.0)	59.4	(29.0)	52.3	(26.5)
DuPont	86–01	55.7	(26.0)	59.3*	(28.8)	51.9	(22.0)
Bank One[7]	86–01	55.3	(30.7)	61.1*	(31.1)	48.9	(29.0)
Siff, Oakley, Marks, Inc.	86–01	54.8	(27.5)	60.9*	(26.4)	48.2	(27.2)
Charles Reeder	86–99	54.6	(29.0)	54.7	(32.1)	54.4	(25.6)
Bear Stearns & Company, Inc.	97–01	54.0	(31.9)	54.5	(32.9)	53.1	(30.7)
Standard & Poor's	94–01	54.0	(28.4)	61.5	(30.4)	45.5	(23.2)
Prudential Financial[8]	86–01	54.0	(25.7)	55.7	(28.1)	52.1	(22.7)
Fannie Mae	98–01	53.3	(26.5)	59.7	(28.5)	44.8	(21.2)
U.S. Chamber of Commerce	86–01	51.9	(26.7)	54.8	(27.5)	48.7	(25.6)
Sears, Roebuck and Company	86–95	51.7	(28.3)	56.8	(30.3)	46.5	(25.1)
Motorola	96–01	51.0	(28.6)	58.5	(31.1)	42.2	(22.7)
UCLA Business Forecast	86–01	49.9	(28.8)	50.9	(31.5)	48.8	(25.8)
Wachovia Securities[9]	96–01	49.8	(25.0)	53.9	(27.5)	44.7	(20.5)
General Motors Corporation	92–01	49.1	(26.0)	47.1	(29.8)	51.3	(21.1)
Comerica[10]	90–01	48.9	(25.6)	49.0	(30.4)	48.8	(19.3)
Goldman Sachs & Company	98–01	47.8	(29.4)	63.8	(25.9)	25.5**	(16.9)
Econoclast	86–01	47.5	(27.0)	45.6	(31.0)	49.5	(21.9)
Prudential Securities[11]	86–96, 00–01	46.6	(31.6)	46.8	(34.0)	46.2	(27.6)
Conference Board	86–01	46.4	(29.7)	53.8	(32.1)	38.4*	(24.5)
Turning Points (Micrometrics)	89–01	46.0	(27.8)	44.2	(30.6)	47.8	(24.4)
Eaton	94–01	45.4	(27.6)	40.9	(28.8)	50.5	(25.4)
JPMorgan Chase[12]	96–01	44.3	(27.9)	52.8	(29.4)	34.3*	(22.3)

	Years in Survey[1]	Average Score		Current-Year Average Score		Next-Year Average Score	
Cahners Publishing Company	86–98	43.0	(25.2)	47.2	(27.8)	38.7*	(21.6)
DRI-WEFA[13]	98–01	42.2	(28.8)	51.7	(29.3)	29.0*	(22.3)
Fairmodel Economica, Inc.	86–93	41.9	(31.5)	45.9	(33.6)	37.9	(28.9)
Chemical Banking[14]	86–95	41.6	(28.9)	45.4	(29.1)	37.1*	(28.1)
Kellner Economic Advisers	97–01	40.9	(20.6)	44.0	(23.7)	37.0	(15.1)
Weyerhaeuser Company	94–00	40.3	(25.1)	42.7	(29.1)	37.8	(20.2)
C.J. Lawrence, Inc.	91–96	39.7	(28.6)	27.9**	(26.0)	56.3	(23.5)
Polyconomics	86–89	38.7	(27.2)	39.7	(29.2)	37.7	(25.4)
Genetski Financial Advisors	92–95, 01	38.3	(30.4)	53.1	(31.3)	21.4**	(18.1)
Morris Cohen & Associates	86–96	37.4*	(28.9)	22.6***	(24.1)	53.7	(24.8)
Bostian Economic Research	86–97	37.0*	(28.6)	26.1***	(28.5)	47.9	(24.2)
Arnhold & S. Bleichroeder	86–93	36.8*	(32.8)	28.5**	(29.5)	46.4	(34.0)
Ford Motor Company	96–01	36.7	(27.7)	37.8	(28.9)	35.0	(26.2)
Inforum–University of Maryland	86–01	36.6**	(26.6)	33.6**	(27.1)	39.8*	(25.8)
Deutsche Banc Alex. Brown[15]	96–01	36.5	(33.1)	37.8	(30.5)	34.6*	(37.0)
Econoviews International, Inc.	86–92	36.3	(28.5)	35.0*	(30.4)	37.7	(26.7)
Morgan Stanley	97–01	33.5	(27.1)	34.4	(29.9)	31.3*	(19.2)
Business Economics, Inc.	86–89	14.3***	(16.5)	13.7***	(16.6)	14.9***	(16.4)

Note: The table shows the average score of forecasters with at least four years of data—seventy forecasters out of a total sample of one hundred four. Numbers in parentheses are standard deviations. *, **, and *** represent significance at the 90 percent, 95 percent, and 99 percent confidence levels, respectively.

1. Years in which there were at least four monthly forecasts for the five variables evaluated in this article.
2. Prior to 07/96, forecasts were from Meyer & Associates.
3. Prior to 12/95, forecasts were from Shawmut National Corporation.
4. Prior to 09/01, forecasts were from Wells Fargo and before 06/96 were from First Interstate Bancorp.
5. Prior to 01/00, forecasts were from National City Bank of Cleveland.
6. Prior to 09/01, forecasts were from Chrysler Corporation.
7. Prior to 11/98, forecasts were from First National Bank of Chicago.
8. Prior to 08/01, forecasts were from Prudential Insurance.
9. Prior to 11/01, forecasts were from First Union Corporation.
10. Prior to 08/92, forecasts were from Manufacturers National Bank of Detroit.
11. Prior to 01/92, forecasts were from Prudential Bache Securities.
12. Prior to 09/01, forecasts were from JPMorgan.
13. Prior to 09/01, forecasts were from WEFA Group.
14. Prior to 02/92, forecasts were from Manufacturers Hanover Trust.
15. Prior to 09/01, forecasts were from Deutsche Morgan Grenfell.

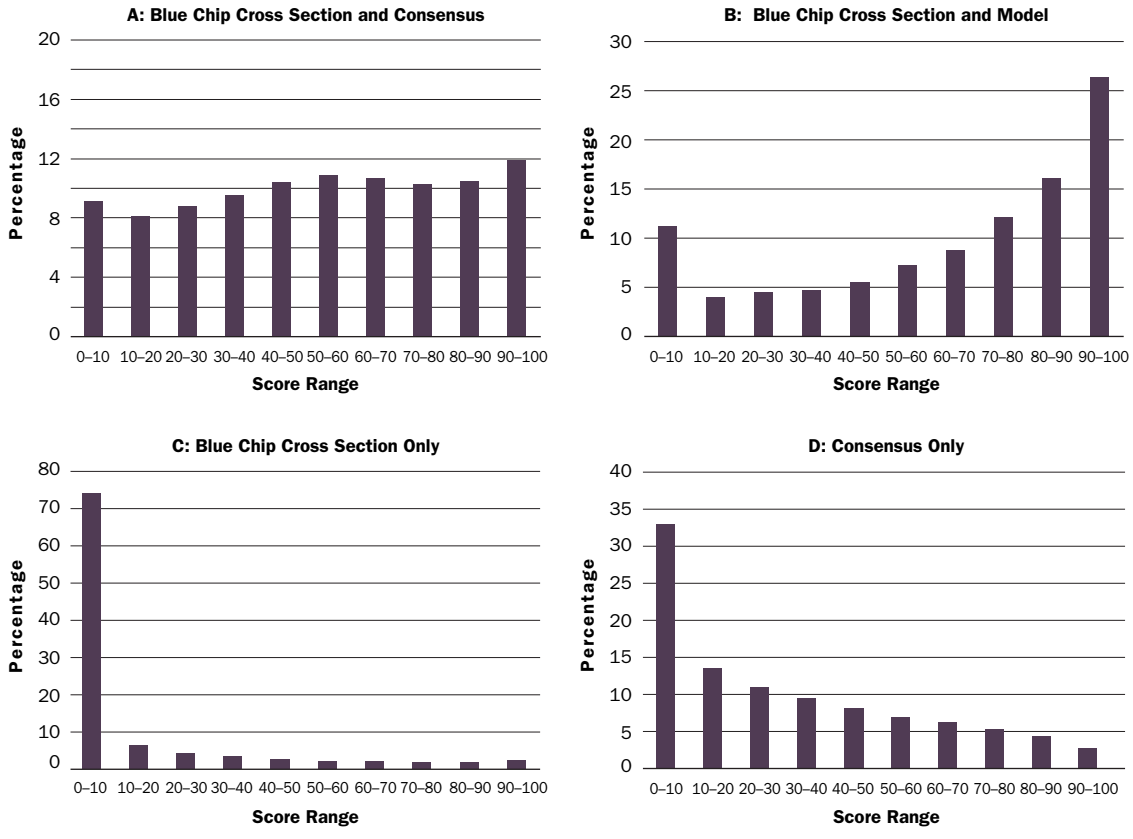
be available. The appropriate choice of a benchmark is complicated by the fact that some series are revised over time. For example, GDP is reported quarterly and revised twice. The advance number is reported in the first month following the end of the quarter, the revised preliminary number is released the next month, and the final number appears three months after the end of the quarter. Also, every July additional revisions may be made to past data. In addition, changes in the definitions of these series may be made. For example, in January 1996 the Bureau of Economic Analysis changed measurement of GDP to a chain-weighted system. This change could be responsible for some of the poor forecasting results observed at the end of 1995 because the forecasts made in 1994 and 1995 for GDP growth over 1995 would be based on the non-chain-weighted series while the GDP data available to assess them would use the chain-weighted numbers. These issues make it important to use vintage data when accessing the accuracy of past forecasts. Vintage, or real-time, data are the data available to the forecaster at a specific time. For instance, vintage January 1990 data are the data that were available to a forecaster at the end of January 1990. For a revised series such as GDP, vintage data would be the advance number for the last quarter of 1989 and the final number for previous quarters. The series used to evaluate forecast accuracy are described in detail in the appendix.

Variance Estimates

As the discussion of methodology showed, if the distribution of the forecast errors is approximately normal, then over time the scores would be

FIGURE 2

Blue Chip Forecast Scores



approximately uniformly distributed. Conversely, an approximately uniform distribution of the scores would be evidence that the underlying assumptions were not grossly violated. This uniformity is important if one is to take seriously the interpretation of the score as the percentage of forecasts expected to be worse than the given one. The uniformity of the distribution of scores will also be sensitive to the choice of estimate of Ω.

Figure 2 plots a histogram of the scores for both the current and next year. Four different variance matrices are used. In panel A the variance is the sum of the cross-sectional variance of the Blue Chip Survey and the variance of the forecast error of the Blue Chip Consensus Forecast. This is the baseline case. In panel B, the variance is the sum of the cross-sectional variance of the Blue Chip Survey and the estimate of the forecast error variance from the theoretical model used in Eisenbeis, Waggoner, and Zha (2002). In panel C only the cross-sectional variance of the Blue Chip Survey is used, and in panel D only the variance of the forecast error of the Blue Chip Consensus is used.

The plot in panel A is virtually uniform, as desired. The histogram in panel B is less uniform and skewed toward higher scores, indicating that the forecast error variance from the model may be too large.[7] Both panels C and D are highly skewed toward lower scores, with the histogram associated with the cross-sectional variance of the Blue Chip Survey in panel C the more skewed. This skewness indicates that individually neither of these matrices captures all of the forecast error variance and that the cross-sectional variance is smaller than the variance of the consensus forecast error. This result is consistent with that of Zarnowitz and Lambros (1987), which showed that in the univariate case the cross-sectional variance underestimates the overall uncertainty of forecasts. It is often claimed that professional forecasters are looking over each other's shoulders and thus produce similar forecasts. The results here are consistent with this view, but they are also consistent with the view that forecasters are all making forecasts close to some hypothetical best forecast but that this best forecast may not be that close to the realized values.

TABLE 2

Score Distribution by Forecast Horizon

	Count	Percentage of Forecast Scores in Each Range									
		0–10	10–20	20–30	30–40	40–50	50–60	60–70	70–80	80–90	90–100
					Current Year						
December	752	16.1	4.8	2.7	3.9	5.1	6.6	4.1	7.6	10.2	39.0
November	750	16.8	4.5	3.5	4.3	4.8	7.1	10.7	9.1	9.5	29.9
October	754	15.5	5.0	4.5	4.0	5.4	5.2	6.4	9.5	17.1	27.3
September	752	13.8	6.4	6.6	5.5	7.0	6.5	6.4	10.1	13.8	23.8
August	745	12.2	7.0	7.5	5.8	9.3	7.7	7.5	9.4	13.4	20.3
July	760	11.1	9.3	7.4	8.7	8.3	10.3	9.3	8.7	12.2	14.7
June	749	9.3	8.3	8.4	8.1	7.7	8.7	10.7	13.5	10.5	14.7
May	743	9.3	7.0	8.9	9.4	9.4	9.0	13.2	11.3	11.0	11.4
April	754	7.6	8.8	8.6	9.2	10.9	11.0	12.3	11.4	12.1	8.2
March	752	8.0	8.2	10.9	10.5	11.2	13.7	10.8	10.9	8.9	6.9
February	739	7.2	9.2	10.1	14.6	10.7	13.3	10.6	8.7	9.2	6.5
January	734	9.5	7.2	8.2	9.5	13.1	12.5	13.6	10.9	9.4	6.0
					Next Year						
December	703	8.7	9.5	7.4	10.1	13.8	12.2	11.0	12.1	9.7	5.5
November	698	7.7	8.7	8.0	9.9	11.5	14.2	12.9	12.6	8.9	5.6
October	705	7.1	8.7	8.1	11.9	10.8	13.9	12.6	11.1	11.1	4.8
September	698	6.2	10.6	8.5	11.3	11.7	11.6	12.3	10.7	10.9	6.2
August	694	6.8	8.8	9.5	9.1	11.4	13.8	13.4	10.1	11.2	5.9
July	699	6.3	9.3	10.9	10.6	11.0	13.9	11.4	10.0	9.7	6.9
June	687	6.3	8.3	10.6	13.0	12.2	11.2	11.2	10.2	11.5	5.5
May	663	5.6	7.8	13.3	11.5	12.1	11.8	13.4	11.0	8.9	4.7
April	667	5.1	9.1	12.0	12.3	12.9	12.4	12.3	9.6	8.7	5.5
March	641	5.8	8.7	11.7	12.9	13.3	13.4	10.6	9.2	8.6	5.8
February	599	5.8	10.9	11.9	13.9	15.7	10.7	10.4	8.8	6.5	5.5
January	553	7.4	9.6	14.5	11.2	14.5	13.2	9.0	9.8	4.9	6.0
All	16,991	9.1	8.1	8.7	9.5	10.4	10.9	10.6	10.3	10.5	11.9

Note: Count is the total number of forecasts in the sample for each month and forecast year. Under the assumptions in this article, all percentages should be approximately equal to 10.

Table 2 shows the distributions for the twenty-four forecast horizons using the baseline estimate for the variance. The last row of this table gives the percentages for the histogram presented in panel A of Figure 2. The other rows can be interpreted as histograms for each forecast horizon. For longer horizons, the scores are approximately uniformly distributed, but for shorter horizons, August through December of the current year, the distribution appears to be U-shaped. A possible explanation for this distribution is that, as the forecast horizon decreases, firms spend fewer resources on the current-year forecast and more on the next-year forecast. In December clients are probably more interested in accurate forecasts of the next calendar year than they are in forecasts of the year that is almost completed. This explanation would account for the higher-than-expected frequency of scores in the lowest range. In turn this pattern would increase and distort the estimate of the variance, which would improve the scores of firms that have good short-term forecasts and explain the higher-than-expected frequency of scores in the high range. Table 2 indicates that this methodology works better for the longer-horizon Blue Chip forecasts.

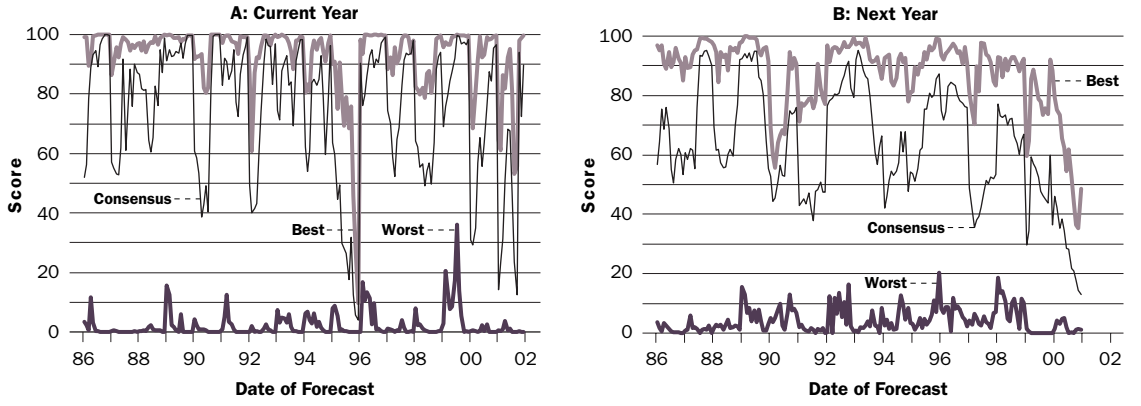
Forecast Performance

Figure 3 plots the score for the consensus forecast for both the current year (panel A) and the

7. In some sense this pattern should be expected since the forecast error variance from the model may be a better proxy for all of Ω instead of just Ω^H. In cases when a too-short time series makes it impossible to use the consensus forecast error variance as a proxy for Ω^H, it may be better to scale the estimate of Ω^H from the model.

FIGURE 3

Blue Chip Consensus Forecast Scores



next year (panel B). The highest and lowest scores for each month are also plotted for comparison. Though the consensus scores vary considerably from month to month, most of the values are above 50 percent. In fact, the average score for the consensus was 75 percent for the current year and 64 percent for the next year. This result means that the consensus forecast was on average more accurate than 75 percent of the current-year forecasts and 64 percent of the next-year forecasts. The three notable exceptions are the current-year forecasts made toward the end of 1995 and the both the current-year and next-year forecasts for 2001 (made in 2001 and 2000, respectively). The low current-year scores toward the end of 1995 are mostly a result of errors in the forecast of GDP, which, as mentioned previously, may stem from the change to a chain-weighted measurement of GDP in 1996. Because forecasts were based on 1995 numbers but all the numbers necessary to evaluate the 1995 forecasts were not available until January 1996, a bias may have been introduced because forecasters were not certain how to adjust their forecasts for the difference in the GDP measure being forecast. This bias is not evident in longer-term forecasts—perhaps because it was small relative to the longer-term forecast errors.

In the forecasts of 2001 made in 2000 (Figure 3, panel B), the forecast errors were large for all the variables except the CPI, with the largest errors occurring in short-term interest rates and GDP. Unlike the 1995 episode, this result can be characterized simply as most of the forecasters having missed the turning point. The 2001 current-year forecasts are a little more complex but are still an interesting case study. The low scores early in the year were caused mainly by large errors in short-term interest rates and GDP, similar to the forecasts of 2001 made in 2000. Early in the year, forecasts were revised and the scores improved. The unforeseen terrorist attack on September 11 caused the economy to be weaker than expected in the last quarter of 2001. This weakness certainly affected the scores of all forecasts, but the largest effect was in current-year forecasts made during the third quarter of 2001. After September 11, forecasts were again revised and were then relatively accurate. This scenario clearly illustrates how economic shocks can cause large swings in the forecast performance. A shock causes prior forecasts to be more inaccurate than they would otherwise be and results in significant revisions, which improve subsequent forecasts.

Table 1 presents the average score of those forecasters with at least four years of data. This criterion leaves seventy forecasters out of one hundred four forecasters in the total sample. Interestingly, out of these seventy forecasters the Blue Chip Consensus Forecast has the highest average score though the average score of several forecasters is almost as good. This result is consistent with the claim that the consensus forecast is a proxy for the hypothetically best forecast and is an argument for giving more weight to the consensus score than to the forecast of any one forecaster.[8]

To further interpret Table 1, note that if a forecaster has average skill, then the mean of T independent scores will be approximately normal with a mean of 50 and a standard deviation of $100/\sqrt{12T}$.[9] For a firm that has been in the sample the entire sixteen years, the average score could be computed on the basis of as many as 372 observations. These observations are not independent because one month's score will be highly correlated with the next month's score. However, forecasts of different

years should be approximately independent. Thus, using T as the number of years a firm has been in the sample gives a more plausible estimate for the standard deviation of the average score reported in Table 1. At the 95 percent confidence level, scores that are more than 1.7 standard deviations apart can be considered statistically different. Putting all of these factors together reveals that, for firms that have been in the sample for the entire sixteen years, scores that are more than 13 percentage points apart can be considered statistically different. Also interesting is the fact that forty-one of the seventy forecasters have average scores that are better than 50 percent, and some of these forecasters have quite long forecasting histories. These figures show that many forecasters have performed consistently well. Conversely, some of the forecasters have scores well below 50 percent and have consistently underperformed.[10]

Table 3 presents the average rank of those forecasters with at least four years of data. Though the exact number of forecasters changes from month to month, the average number of forecasters is approximately forty-seven, and in almost all months there were between forty-two and fifty-two forecasters. For this reason, it is not necessary to scale the ranks to some common interval before averaging. Again, the consensus forecast has the best average rank although several forecasters are close. However, the standard deviation of the consensus forecast rank is less than half that of the others. This result implies that the consensus is consistently among the best forecasts even when its score is relatively low. Figure 4 illustrates this finding. The consensus forecast rank is plotted with the Macroeconomic Advisors' forecast rank for both the current- and next-year forecasts. These figures show how much more volatile the Macroeconomic Advisors' ranks are as compared to the consensus forecast ranks. The plots for all the other top-ranked forecasters are similar.

Improving the Consensus

Instead of using the consensus forecast, would it be better to form a "superconsensus" using only highly ranked forecasters? Table 4 shows the results from using the best forecasters from recent years. The table groups the results of the best one, three, five, ten, fifteen, and twenty-five forecasters for periods from one to five years. The performance of the superconsensus can be compared to that of the regular consensus. Only data available at the time the superconsensus is formed are used in its construction. Panel A compares the average scores of various superconsensuses, and panel B compares the average ranks. These results imply that there is at best a very small gain in average score or rank in using a superconsensus forecast, but there is an increase in the standard deviation of the rank. More interesting is the observation that if only a few forecasters are used, then it is clearly best to use those with a long track record of superior forecasts. However, if more than five forecasts are averaged, there seems to be little advantage to using more than the prior two years to select the best forecasters.

The results in this article indicate that forecasters are all making forecasts close to some hypothetical best forecast but that this best forecast may not be that close to the realized values.

Conclusions

A consistent evaluation of forecasts over time that also respects their multivariate character is essential if the forecasts are to be used for decision making. Having both a cross section and a time series of forecasts, as in the Blue Chip Survey, gives one the ability to perform such an evaluation. The methodology developed in Eisenbeis, Waggoner, and Zha (2002) gives consistent results for the Blue Chip Survey Forecasts, particularly at longer forecast horizons. Furthermore, the methodology reveals that the Blue Chip Consensus Forecast consistently performs better than any of the individual forecasters do. This result is a "reverse Lake Wobegon" effect: none of the forecasters are better than the average forecaster. While no forecaster had a higher average score than the consensus forecast, several were indistinguishably close, and many had average scores well above 50 percent. There are superior forecasters, but no individual has access to all of the independent information from all of the forecasts that is incorporated into the consensus forecast.

8. The result is also consistent with the Ottaviani and Sorensen (2003) hypothesis that the forecasts are unbiased.
9. The mean of a uniform random variable on (0, 100) is 50, and its standard deviation is $100/\sqrt{12}$. The standard deviation of the mean of T independent random variables, each with standard deviation σ, is $\sigma/\sqrt{T}$.
10. This pattern also suggests that there may be some survivorship effects in the data.

TABLE 3

Average Rank

	Years in Survey	Average Rank		Current-Year Average Rank		Next-Year Average Rank	
BC Consensus	86–01	13.2***	(4.8)	12.4***	(5.1)	13.9***	(4.3)
Moody's Investors Service	98–01	13.5*	(10.3)	16.4	(11.0)	9.4**	(7.5)
Security Pacific National Bank	86–92	14.3**	(10.9)	14.4**	(11.6)	14.3**	(10.3)
NationsBank	93–98	15.1**	(12.9)	16.3*	(12.8)	13.8**	(13.0)
Mortgage Banker Assn. of America	86–01	15.5***	(10.4)	14.3***	(9.9)	16.7**	(10.7)
Macroeconomic Advisors	86–01	15.5***	(10.8)	13.7***	(9.9)	17.3**	(11.5)
Northern Trust Company	86–01	17.9**	(12.0)	18.3**	(11.6)	17.5**	(12.4)
Bank of America	87–01	18.0**	(11.5)	19.3*	(12.4)	16.6**	(10.3)
U.S. Trust Company	86–01	18.5**	(12.7)	17.6**	(12.0)	20.0*	(13.6)
CoreStates Financial Corporation	88–98	19.2*	(12.3)	22.5	(13.1)	15.8**	(10.5)
Peter L. Bernstein, Inc.	86–89	19.6	(12.6)	20.5	(12.8)	18.8	(12.4)
Equitable Life Assurance	86–91	19.9	(11.3)	17.8	(11.7)	23.0	(10.0)
Wayne Hummer Investments, LLC	86–01	20.2*	(11.3)	21.9	(12.3)	18.3**	(9.9)
Chicago Capital, Inc.	96–00	20.5	(15.5)	21.0	(13.8)	20.1	(17.2)
Fannie Mae	98–01	20.9	(10.9)	20.8	(11.2)	21.1	(10.7)
Pennzoil Company	86–89, 92–93	21.0	(10.9)	19.7	(12.1)	22.2	(9.5)
Merrill Lynch	86–01	21.0	(12.4)	21.1	(12.3)	20.9	(12.4)
Dean Witter Reynolds & Company	86–91	21.0	(13.1)	22.1	(11.7)	19.6	(14.6)
Wells Capital Management	91–01	21.3	(12.7)	20.7	(11.3)	22.1	(14.2)
Georgia State University	86–01	21.4	(12.7)	23.1	(13.0)	19.6*	(12.2)
National Association of Home Builders	90–01	21.6	(10.8)	21.9	(11.1)	21.2	(10.5)
PNC Financial Corporation	88–98	21.7	(11.1)	23.7	(11.5)	19.6	(10.3)
DaimlerChrysler AG	86–01	21.9	(12.7)	20.5	(12.3)	23.3	(13.0)
Bear Stearns & Company, Inc.	97–01	22.2	(17.1)	24.3	(16.9)	18.8	(17.3)
La Salle National Bank	86–91, 97–01	22.3	(12.3)	21.6	(12.8)	23.2	(11.7)
National City Corporation	86–01	22.4	(11.8)	23.3	(12.3)	21.4	(11.1)
Fleet Financial Group	91–99	22.4	(11.8)	22.0	(12.8)	22.8	(10.7)
Eggert Economic Enterprises, Inc.	86–01	22.4	(12.1)	25.6	(12.0)	19.1*	(11.4)
Evans Group	86–01	22.4	(13.6)	20.7	(13.2)	24.3	(13.7)
Metropolitan Life Insurance Company	86–96	22.6	(11.0)	23.3	(12.8)	21.9	(8.8)
DuPont	86–01	22.8	(11.8)	23.4	(12.1)	22.2	(11.5)
University of Michigan M.Q.E.M.	86–96	22.8	(12.3)	19.3*	(11.7)	26.2	(12.0)
Standard & Poor's	94–01	22.9	(13.6)	19.9	(13.5)	26.3	(12.8)
Dun & Bradstreet	89–99	23.2	(13.9)	22.7	(14.4)	23.8	(13.5)
Bank One	86–01	23.2	(14.9)	21.4	(14.3)	25.1	(15.4)
Wachovia Securities	96–01	23.3	(14.0)	24.9	(14.7)	21.4	(13.0)
Siff, Oakley, Marks, Inc.	86–01	23.5	(12.8)	23.3	(12.0)	23.7	(13.6)
Chase Manhattan Bank	88–00	23.7	(14.3)	22.8	(13.5)	24.8	(15.2)
Prudential Insurance	86–01	23.8	(12.7)	26.0	(12.5)	21.3	(12.5)
Charles Reeder	86–99	23.8	(14.9)	25.7	(15.3)	21.8	(14.2)
Goldman Sachs & Company	98–01	24.5	(14.6)	16.8	(13.0)	35.2*	(8.8)
U.S. Chamber of Commerce	86–01	25.3	(11.7)	26.5	(11.1)	23.9	(12.3)
Sears, Roebuck and Company	86–95	25.7	(12.4)	24.5	(13.0)	26.9	(11.8)
Motorola	96–01	25.9	(12.2)	24.7	(13.3)	27.2	(10.8)
Comerica	90–01	26.1	(13.4)	28.3	(13.4)	23.7	(13.0)
UCLA Business Forecast	86–01	26.2	(13.7)	28.0	(13.7)	24.4	(13.5)
General Motors Corporation	92–01	27.0	(12.9)	29.3	(13.9)	24.4	(11.2)
Prudential Securities	86–96, 00–01	27.2	(13.7)	26.5	(14.7)	28.3	(11.9)

	Years in Survey	Average Rank		Current-Year Average Rank		Next-Year Average Rank	
Econoclast	86–01	27.7	(12.7)	30.5*	(12.7)	24.7	(12.0)
Turning Points (Micrometrics)	89–01	27.7	(12.9)	30.6*	(12.8)	24.6	(12.3)
Eaton	94–01	27.7	(14.0)	32.8*	(11.9)	22.0	(14.0)
Conference Board	86–01	29.2	(13.3)	27.2	(13.8)	31.3**	(12.5)
Kellner Economic Advisers	97–01	29.3	(11.9)	31.5	(10.7)	26.4	(12.9)
DRI-WEFA	98–01	29.8	(12.9)	27.8	(13.1)	32.5	(12.4)
JPMorgan Chase	96–01	29.9	(13.2)	26.1	(15.3)	34.4*	(8.4)
Fairmodel Economica, Inc.	86–93	30.4	(13.9)	30.2	(14.2)	30.6	(13.7)
C.J. Lawrence, Inc.	91–96	30.5	(14.7)	34.4*	(13.9)	25.0	(14.1)
Arnhold & S. Bleichroeder	86–93	30.8	(15.6)	35.8**	(12.3)	25.1	(17.1)
Cahners Publishing Company	86–98	30.9*	(10.9)	30.2*	(11.4)	31.7**	(10.3)
Polyconomics	86–89	31.0	(13.0)	31.9	(12.8)	30.2	(13.3)
Chemical Banking	86–95	31.1*	(13.1)	29.9	(12.0)	32.5*	(14.2)
Bostian Economic Research	86–97	31.1*	(14.9)	35.6***	(14.6)	26.6	(13.8)
Inforum–University of Maryland	86–01	31.1**	(13.5)	33.9***	(11.9)	28.0	(14.5)
Genetski Financial Advisors	92–95, 01	31.2	(13.9)	24.9	(14.0)	38.4**	(9.7)
Weyerhaeuser Company	94–00	31.8	(12.9)	31.4	(13.5)	32.3*	(12.5)
Econoviews International, Inc.	86–92	31.9	(11.1)	33.9*	(10.9)	30.0	(11.1)
Deutsche Banc	96–01	31.9	(18.1)	31.9	(17.7)	32.0	(18.9)
Morris Cohen & Associates	86–96	31.9*	(12.6)	38.4***	(9.9)	24.6	(11.3)
Morgan Stanley	97–01	34.0*	(14.5)	34.1*	(14.6)	33.7*	(14.5)
Ford Motor Company	96–01	34.2*	(13.1)	35.0**	(12.4)	33.1*	(14.3)
Business Economics, Inc.	86–89	40.7**	(6.7)	41.4**	(5.0)	39.9**	(8.1)

Note: Numbers in parentheses are standard deviations. *, **, and *** represent significance at the 90 percent, 95 percent, and 99 percent confidence levels, respectively.

FIGURE 4

Macroeconomic Advisors and Blue Chip Consensus Forecast Ranks

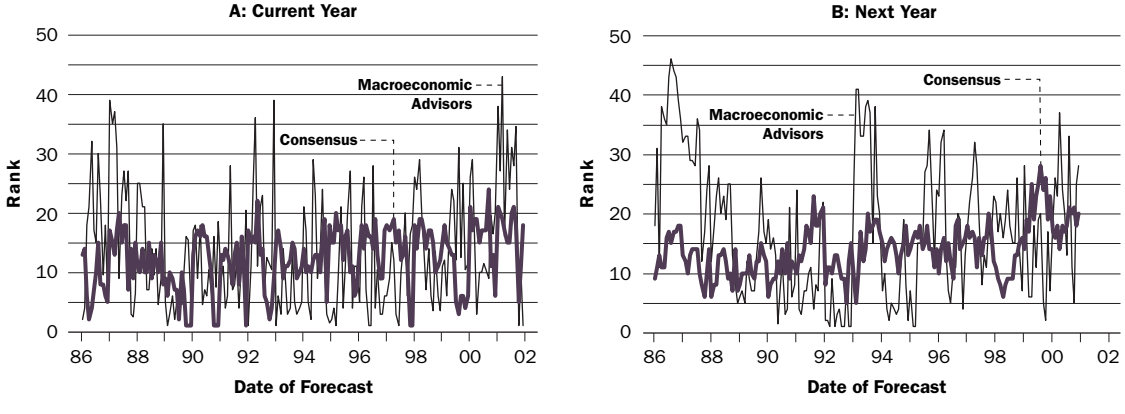


TABLE 4

Average of "Super Consensus" Scores and Ranks, 1992–2001

	Over Prior Year	Over Prior 2 Years	Over Prior 3 Years	Over Prior 4 Years	Over Prior 5 Years
			Scores		
Best forecaster	54.7 (31.1)	58.9 (28.5)	63.9 (26.1)	63.8 (26.8)	62.3 (27.0)
3 best forecasters	60.4 (27.1)	65.4 (25.2)	66.0 (24.6)	66.2 (24.8)	65.6 (23.9)
5 best forecasters	63.6 (25.3)	66.7 (23.9)	66.7 (23.5)	66.6 (23.3)	66.7 (23.3)
10 best forecasters	66.5 (23.6)	67.4 (22.9)	66.9 (22.8)	67.3 (23.5)	67.2 (22.5)
15 best forecasters	67.0 (23.3)	67.4 (23.3)	67.2 (23.0)	67.1 (23.2)	67.1 (22.7)
25 best forecasters	66.8 (23.0)	66.8 (22.8)	67.2 (22.9)	66.9 (22.9)	67.0 (22.5)
Consensus forecast	66.3 (23.0)	66.3 (23.0)	66.3 (23.0)	66.3 (23.0)	66.3 (23.0)
			Ranks		
Best forecaster	21.8 (15.5)	20.5 (13.9)	17.1 (11.6)	17.3 (11.7)	17.4 (11.5)
3 best forecasters	18.6 (12.8)	15.7 (9.7)	15.3 (9.8)	14.9 (9.2)	15.0 (8.3)
5 best forecasters	16.7 (10.8)	14.3 (8.1)	14.4 (8.9)	14.2 (7.5)	13.9 (7.3)
10 best forecasters	14.1 (7.4)	13.4 (6.0)	13.9 (6.4)	13.5 (6.0)	13.7 (5.8)
15 best forecasters	13.7 (6.4)	13.3 (5.5)	13.6 (5.3)	13.7 (5.3)	13.8 (5.2)
25 best forecasters	13.9 (5.4)	13.8 (4.7)	13.4 (4.7)	13.5 (4.8)	13.7 (4.7)
Consensus forecast	14.2 (4.7)	14.2 (4.7)	14.2 (4.7)	14.2 (4.7)	14.2 (4.7)

Note: Standard deviations are in parentheses.

APPENDIX

Data Description

Gross domestic product: 1986–95, not chained; 1996–current, chained 1996 dollars. (Note that data are revised only through March after the forecast year.) Source: U.S. Department of Commerce, Bureau of Economic Analysis, *Gross Domestic Product*, table 3.

Consumer price index: CPI-U is all urban consumers. Source: U.S. Department of Labor, Bureau of Labor Statistics, *Consumer Price Index*.

Unemployment rate: Unemployment rate (all workers). Source: U.S. Department of Labor, Bureau of Labor Statistics, *Employment Situation*.

Three-month Treasury bill: Three-month Treasury bills, secondary market (monthly average). Source: Board of Governors of the Federal Reserve System, "Selected Interest Rates," Release H.15.

Corporate bonds—1986–95: Moody's Corporate Bond Yield, Aaa (monthly average). Source: Moody's Investors Service, Inc.

Ten-year Treasury note—1996–current: Ten-year Treasury note yield at constant maturity (monthly average). Source: Board of Governors of the Federal Reserve System, "Selected Interest Rates," Release H.15.

REFERENCES

Bates, John M., and Clive W.J. Granger. 1969. The combination of forecasts. *Operational Research Quarterly* 20, no. 4:451–68.

Clemen, Robert T. 1989. Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting* 5, no. 4:559–83.

De Menezes, Lilian M., Derek W. Bunn, and James Taylor. 2000. Review of guidelines for the use of combined forecasts. *European Journal of Operational Research* 120, no. 1:190–204.

Eisenbeis, Robert, Daniel Waggoner, and Tao Zha. 2002. Evaluating *Wall Street Journal* survey forecasters: A multivariate approach. *Business Economics* 37 (July): 11–21.

Newbold, Paul, and Clive W.J. Granger. 1974. Experience with forecasting univariate time series and the combination of forecasts. *Journal of the Royal Statistical Society*, ser. A, 137, pt. 2:131–46.

Ottaviani, Marco, and Peter Norman Sorensen. 2003. The strategy of professional forecasting. Unpublished working paper.

Zarnowitz, Victor, and Louis A. Lambros. 1987. Consensus and uncertainty in economic prediction. *Journal of Political Economy* 95 (June): 591–621.

Emotion and Financial Markets

LUCY F. ACKERT, BRYAN K. CHURCH, AND RICHARD DEAVES

Ackert is a professor at the Michael J. Coles College of Business at Kennesaw State University and a visiting scholar at the Atlanta Fed. Church is a professor at the DuPree College of Management at Georgia Tech. Deaves is a professor at the Michael C. DeGroote School of Business at McMaster University. *The authors thank Jerry Dwyer, Mark Fisher, and Larry Wall for helpful comments.*

> We are merely reminding ourselves that human decisions affecting the future, whether personal or political or economic, cannot depend on strict mathematical expectation, since the basis for making such calculations does not exist; and that it is our innate urge to activity which makes the wheels go round, our rational selves choosing between the alternatives as best we are able, calculating where we can, but often falling back for our motive on whim or sentiment or chance.
>
> —John Maynard Keynes (1964, 162–63)

The popular press commonly reports that psychology drives financial decision making and moves asset prices. Yet traditional implementations of financial economic models routinely assume that individuals incorporate information into their decision processes using the rules of probability and statistics with calculated, unemotional logic. This assumption leaves little room for the influence of emotion. Furthermore, when economists have included emotion in describing the behavior of financial markets, emotion is often characterized as causing unwarranted and undesirable price movements. For example, in his book *Irrational Exuberance*, Robert Shiller states that investors' emotional state "is no doubt one of the most important factors causing the bull market" recently experienced in the United States (2000, 57).

Is a "rational" person a cool, unemotional user of logic and the laws of probability? Two characters from the popular television and movie series *Star Trek* provide an answer. Mr. Spock—who is half Vulcan, a species that suppresses emotion and prizes logic—is presented as a rational thinker who thoroughly considers every piece of information. In contrast, Captain Kirk is likely to respond emotionally. Yet Kirk is portrayed as a good decision maker. Though Spock fully analyzes each situation, he gets too caught up in the details. Emotion allows Kirk to focus and enhances his ability to make critical decisions.

A vast psychological literature shows that emotional state can significantly affect decision making (Elster 1998; Hermalin and Isen 2000). In contrast to studies by some other financial economists, this article demonstrates that emotion actually enhances an individual's ability to make rational choices (see also Frank 1988; Damasio 1994; LeDoux 1996; Elster 1998; Isen 1999). Emotion allows people to transcend the details, prioritize, and focus on the decision to be made. Emotion can drive behavior that is consistent with economic predictions.

Understanding what behavior is economically rational is complex. Behavioral research in finance applies lessons from psychology to financial decision making. One aspect of individual psychology that has received a considerable amount of attention is that of cognitive limitations. Individuals are limited in their abilities to encode, process, and retrieve information. In some cases, psychologists argue that these limitations result in biased judgments. Psychologists posit that individuals develop rules of thumb, or heuristics, to promote good decision making with minimal processing. Heuristics allow people to make decisions while economizing on processing. Although individuals develop habits that often serve them well, these habits might occasionally lead them astray. Behavioral finance research has focused primarily on these biases, paying less attention to the role of emotion.

The examination of cognitive aspects of financial behavior in isolation is troublesome and may be

misleading. Emotional reactions or evaluations occur at a very early stage and are more basic than cognitive evaluations (Zajonc 1980; LeDoux 1996). Perceptions encompass emotional aspects, which subsequently guide judgment and decision making. Furthermore, theorists recognize that emotion and cognition are interdependent, rather than competing, influences (Simon 1967).

The purpose of this article is to provide a framework from which future research on emotion in financial markets can build.[1] The discussion begins by describing the vastly different views of human behavior held by economists and psychologists. After differentiating their approaches, the article defines the term *emotion*, describes how emotions can be categorized, and then describes how emotions influence human behavior. The focus then turns to three particular aspects of emotion and financial decision making: emotional disposition and stock market pricing, the feeling of regret, and investors' emotional response to information. The conclusion considers emotion and the traditional financial economics paradigm.

Individual psychology plays a limited role in finance theory, which assumes that individuals maximize expected utility, with expectations derived using the rules of probability and statistics.

The Psychology of Economists

Economists and psychologists take strikingly different approaches to the study of human decision making. Some have cultivated dialogues across the disciplines, but the success of a discussion is dependent on the particular issue (Hogarth and Reder 1986). Economists argue that some empirical analyses in psychology lack a theoretical basis. Furthermore, economists argue that empirical evidence provided by psychologists gives little insight into people's decisions because these studies fail to provide participants with meaningful (for example, monetary) incentives and lack the discipline that markets give to behavior. Some psychologists, on the other hand, argue that economists' models bear little relation to actual behavior. Although economists typically assume that individuals choose among alternatives in an internally consistent way, behavior in detailed experiments is often inconsistent with this assumption.

Individual psychology plays a limited role in finance theory, which assumes that individuals maximize expected utility, with expectations derived using the rules of probability and statistics. The efficient market hypothesis (EMH), though certainly not the only economic model describing financial market behavior, has been the central paradigm in financial economics for more than thirty years. In an efficient market, as Fama (1991) defined it, prices reflect all available information. According to this hypothesis, prices should reflect information in such a way that the marginal profit acquired by acting on information does not exceed the marginal cost of acquiring the information. This simple model revolutionized prevailing thought in the 1970s on how markets function. Early empirical evidence supported the EMH. In his first review of an already vast body of evidence, Fama proclaimed that the "support of the efficient markets model is extensive, and (somewhat uniquely in economics) contradictory evidence is sparse" (1970, 416). In fact, Michael Jensen, another prominent scholar, asserted that "there is no other proposition in economics which has more solid empirical evidence supporting it than the efficient market hypothesis" (1978, 95).

More recently, departures from the predictions of the EMH have been reported, and many now argue that markets do not efficiently incorporate information (Haugen 1995; Shiller 2000; Shleifer 2000). Fischer Black (1986) provides a model in which some investors trade on noise rather than information, and, as a result, market prices are not efficient. Black's noise traders' behavior is not driven by news related to an asset's underlying value and may not be fully rational. Noise traders may trade because they mistakenly believe they are trading on information or perhaps because they simply like to trade. Some empirical evidence is consistent with the proposition that irrational behavior results in market inefficiencies. For instance, Daniel, Hirshleifer, and Teoh (2001) argue that investors frequently make large errors that impede the EMH's corrective forces.

The recognition that individual behavioral influences affect market outcomes initiated a new research stream in financial economics—behavioral finance. Behavioral finance research applies lessons from psychology to financial decision making. This research has focused primarily on cognitive biases, paying scant attention to emotion's role. Emotion is clearly an important aspect of human psychology though it is not fully understood. In fact, no generally accepted definition of emotions exists.

What Are Emotions, and How Do They Influence Behavior?

There is a vast body of research on emotions, but the term is seldom defined. Rather, examples of emotional states are provided. Emotion can be defined loosely as a physiological state of arousal triggered by beliefs about something (Elster 1998). Arnold (1960) defines emotion as "the felt tendency toward anything intuitively appraised as good (beneficial), or away from anything intuitively appraised as bad (harmful)" (182). A strict definition of the term is complex because emotion has cognitive, physiological, social, and behavioral aspects (Solomon 2000).

For many, the substance of an emotion is feeling. But emotions are evaluative rather than purely bodily sensations or cognitive judgments (Frijda 2000). An emotion may have no cognitive basis whatsoever: "a rose smells good because it smells good" (Frijda 2000, 63). Each individual has a personal assessment of whether an object or state is good or bad. Emotions are evaluative in that they evoke positive or negative valences that can be described using bipolar scales that define a continuous spectrum from unpleasantness to pleasantness—for example, unhappy to happy or pessimistic to optimistic (Bradley and Lang 2000, 247).

Despite the lack of a unified definition of emotion, there is some agreement on the set of emotions that exist. According to Elster (1998), some states are clearly emotions, including, for instance, anger, hatred, guilt, regret, fear, pride, elation, joy, and love. Elster further argues that these emotional states can be differentiated from other mental states on the basis of six features put forth long ago. These features do not provide a complete definition of emotion because not even one feature is an element of every emotion. Yet these six features remain central to current discussion and provide a framework for understanding what an emotion is. The brief descriptions that follow use one emotion—regret—for illustrative purposes.

1. **Cognitive antecedents.** Emotions are triggered by beliefs. An investor regrets an investment decision because she believes that bad outcomes could have been avoided.
2. **Intentional objects.** Emotions are about something. The object of an emotion is usually the cognitive antecedent. For example, the poorly performing investment is the object of the regretful investor.
3. **Physiological arousal.** Changes in hormonal conditions and the autonomic nervous system accompany emotions. The regretful investor may feel pangs, a hollow stomach, or depression.
4. **Physiological expressions.** Observable expressions characterize emotions. Facial expressions, posture, voice intonation, and outward appearance are noteworthy. The regretful investor may appear pale, with slumped shoulders.
5. **Valence.** Emotions can be placed on a scale with pleasure at one extreme and pain at the other. Valence, or the experience of pleasure versus pain, translates to happiness or unhappiness. The regretful investor is decidedly unhappy about the poor investment outcome.
6. **Action tendencies.** Emotions are associated with a tendency to act. The regretful investor might take actions to avoid being exposed to similar investment opportunities.

Where Do Emotions Come from, and Where Do They Take Us?

As a result of millions of years of selection, people are well engineered to solve problems repeatedly encountered during evolution. The ability to learn and adapt is critically important to survival. Many emotions are useful responses that result from evolutionary conditioning (Frank 1988; LeDoux 1996). For example, fear is a natural, rational, and useful response in a dangerous situation. In fact, emotional reactions and preferences can form with no conscious recognition of the stimuli (Zajonc 1980). According to Goleman (1995), an individual's success in life depends as critically on what Goleman calls the "emotional quotient" as on the individual's intelligence quotient (IQ). Romer (2000) argues that some behaviors reported to be irrational or inconsistent with well-defined preferences might be better explained by allowing complicated feelings in economic models. People's preferences may be defined by arguments that are not reflected in some economic models.

Path-breaking work by Damasio (1994) indicates that a lack of emotion has striking effects on decision making. Damasio offers behavioral and physiological evidence in support of the hypothesis that decision making is intertwined with emotion. He studied brain-damaged patients who had impaired emotional responses even though they retained their cognitive abilities. The patients were emotionally flat as a result of frontal brain lobe damage, yet

1. This article does not attempt to provide an overview of the vast literature on investor psychology. A recent review is provided by Hirshleifer (2001).

their knowledge, attention, memory, language, and abstract problem solving were unaffected. These individuals had difficulty making decisions and were unable to plan for the future or choose a course of action. Damasio hypothesizes a connection between flawed reason and impaired feelings.

A patient, referred to as Elliot, provides an example. While in his thirties, Elliot had experienced a severe change in personality following the removal of a brain tumor. Before his illness, Elliot was a successful husband, father, businessman, and member of the community. After surgery, Elliot could not hold a job, manage his time, or maintain social relationships. Yet his IQ remained in the superior range. Extensive testing indicated that Elliot's memory, perceptual ability, language, arithmetic ability, and ability to learn new material were unaffected. Elliot had normal intellectual functioning but was completely unable to make a decision, particularly one of a personal or social nature. Elliot himself reported that he no longer responded in the same way to emotional stimuli. What had once caused a strong emotional response now caused no reaction whatsoever. Although he could reason through a problem, he could not choose a course of action. For instance, if given the task of sorting clients' documents, Elliot could easily understand the material. Yet his attention might be easily diverted, or he might spend hours reading one document, or he might just as easily spend an extended period of time pondering whether the classification scheme was appropriate. Not surprisingly, it was not long before his employment was terminated. A series of financially ruinous ventures followed.

A large body of literature supports the theory that positive mood allows individuals to better organize and assimilate information and facilitates creative problem solving.

Damasio concludes that feelings have a very strong influence on reasoning. A complete understanding of human behavior requires recognition of the interconnection between the brain and the body. Reason and emotion are part of the human organism. Although emotional responses typically are characterized as irrational, recent research suggests that emotion and rational decision making are complementary.

Neurobiological studies (Damasio 1994; LeDoux 1996) indicate that emotion improves decision making in two respects. First, emotion pushes individuals to make some decision when making a decision is paramount. In some situations in life, so many options exist that an individual could devote excessive amounts of time to the decision-making process. An individual could simply become overwhelmed by the possibilities. Emotion provides a coping mechanism and allows individuals to focus without being caught up in the details.

Second, emotion can assist in making optimal decisions. A vast psychological literature shows that emotional state can significantly affect decision making (Elster 1998; Hermalin and Isen 2000). While strong emotional responses are often associated with poor decisions (particularly those of a financial nature), recent research in psychology indicates that the absence of emotions can also lead to suboptimal decisions. Emotion helps to optimize over the cost of optimization. Even mild emotional states can affect behavior (Isen 2000). Positive feelings can make it easier to access information in the brain, promote creativity, improve problem solving, enhance negotiation, and build efficient and thorough decision making. Emotion facilitates optimal-choice behavior when a person is provided with several courses of action (Rolls 1999).

Little attention has been paid to the direct role of emotion on choices of a financial nature. Recently, Lo and Repin (2001) studied the physiological characteristics of professional securities traders while they are engaged in live trading. They report significant correlation between market events and physiological characteristics including skin conductance and cardiovascular data. They conclude that emotion is an important determinant of a trader's ability to survive in financial markets. Other recent research has focused on the role of emotion in a more indirect fashion. Specifically, anomalous financial behavior is frequently attributed to emotion. The next section reviews some of these studies.

Emotional Disposition

A person's current emotional state may influence financial decision making. For example, an individual in a good mood because of recent experience or current position in life brings this positive outlook to the task at hand. Ashbury, Isen, and Turken (1999) argue that a positive mood enhances individual performance on many cognitive tasks. A large body of literature supports the theory that positive mood allows individuals to better organize and assimilate information and facilitates creative problem solving.

Others have argued that evidence on the importance of emotional disposition is provided by empirical results at the aggregate level (for instance, Hirshleifer and Shumway 2003; Kamstra, Kramer, and Levi 2003). Using data from twenty-six international stock exchanges, Hirshleifer and Shumway argue that good moods resulting from morning sunshine lead to higher stock returns.[2] The argument is that, because people are more optimistic on a sunny day, they are more inclined to buy stocks.

These aggregate studies of the effect of mood on stock market pricing do not provide evidence on how individual behavior translates into market outcomes. Yet theoretical and experimental evidence suggests that even when individual behavior is, on average, characterized as irrational, market outcomes can be consistent with rational pricing (Ackert and Church 2001; Jamal and Sunder 1996, 2001; Chen and Yeh 2002).

More fundamentally, however, the relationship between mood and risk tolerance is not well established. Risk aversion is important because changes in risk aversion affect how much an individual is willing to pay for a stock in response to changes in mood. When an individual becomes elated, perhaps because of good weather, he or she might become more willing to buy stock at higher prices. If melancholy is associated with greater risk aversion, an individual suffering from depression might associate lower valuations with stocks. The literature does not provide compelling evidence that optimism or euphoria leads to lower risk aversion or that depression or a poor mood leads to increased risk aversion.

According to Thaler and Johnson (1990), it is extremely difficult to make generalizations about preferences toward risk. They conclude that after a series of winning gambles, individuals are willing to take on more risk so that risk aversion declines after prior gains.[3] However, after an initial loss, experimental participants become more risk averse. Other research shows that happy people are more optimistic and assign higher probabilities to positive events (Wright and Bower 1992). Yet decision-making research shows that even though happy people are more optimistic about their likelihood of winning a gamble, they are much less willing to actually take the gamble (Isen, Nygren, and Ashby 1988). They are more risk averse. People in a good mood are less likely to gamble because they do not want to jeopardize their good mood. Thus, it is not clear how positive and negative emotional states affect risk preferences and, in turn, translate into market pricing.

Clearly, clinical depression is quite different from a simple bad mood. Depression has a biochemical basis and can occur without cognitive appraisals. A person with no chemical imbalances might naturally experience anxiety in certain situations (for example, a job interview), but a depressed person might feel chronically anxious with a view that the world is an inexhaustible source of threats. Furthermore, the modern view of depression recognizes that the condition may involve altered brain circuitry (LeDoux 2002).

As with the evidence on the effect of mood on risk choices, experimental evidence concerning the relationship between risk tolerance and depression fails to provide a clear picture. Some researchers question the importance of anxiety and depression in explaining choices across risky alternatives (Hockey et al. 2000). Others conclude that risk aversion is correlated with depressive tendencies (Eisenberg, Baron, and Seligman 1998). Importantly, as these authors recognize, risk aversion is correlated with anxiety and depression.[4] Eisenberg, Baron, and Seligman report that the correlation between depressive symptoms and risk aversion arises from the correlation with anxiety.[5] The fundamental issue remains unresolved. While a depressed person shying away from risk for no apparent reason may appear to be irrational, it may be perfectly rational for an anxious person to move toward safer alternatives. Again, much research needs to be done to move toward a definitive conclusion.

2. Another stream of research in financial economics investigates the impact of investor sentiment on asset pricing. Sentiment is broadly defined as the deviation in asset returns from that predicted by the fundamental determinants of asset value, such as dividends (Lee, Shleifer, and Thaler 1991). The source of the sentiment may be noise (Black 1986). Though not postulated in the literature, the source of sentiment, as discussed in the finance literature, could also be changes in the emotional disposition of the population of investors.
3. Barberis, Huang, and Santos (2001) have formulated a theoretical model of this behavior that predicts that individuals will become more risk averse after a fall in stock prices. Investors derive utility from changes in wealth and are more sensitive to decreases in wealth than to increases.
4. Note that this study, like many others, is based on hypothetical questions, and, thus, decisions are not financially motivated. Furthermore, the measure of depressive symptoms is based on a survey given to a sample of students registered in a college course. The incidence of clinically diagnosed depression in this sample is not reported.
5. Interpretation of the results becomes even more difficult because Raghunathan and Pham (1999) find that anxiety and sadness have distinct influences on behavior.

Regret

Regret is an emotion that colors an investor's current disposition. Some claim that fear of regret can drive certain financial decisions. This emotion is counterfactual in that it is generated by thoughts about what might have happened but did not. Clearly, regret is a negative emotion. An investor may regret a bad investment decision but is not likely to regret a good one.

Psychologists recognize the important impact regret can have on decision making. According to Kahneman and Tversky (1979), individuals have strong desires to avoid the feeling of regret. They argue that a number of the implications of expected utility theory are not corroborated by experimental evidence and provide an alternative to the standard economic paradigm—prospect theory. Central to their theory is the notion of loss aversion: Individuals will change their behavior in order to avoid recognizing losses. Experimental subjects, given the hypothetical choice between $500 with certainty versus a coin flip for $1,000, will usually choose the former: they are risk averse.[6] This risk aversion, however, would also imply that subjects should choose a loss of $500 with certainty rather than the flip of a coin where they can either return to zero or lose an extra $500. In the experiments, however, most subjects choose the gamble: They are risk loving in the domain of losses. Kahneman and Tversky argue that individuals wish to avoid the negative feeling of regret that would occur if they have to recognize a loss, and so they alter their "normal" risk-averse tendencies. The results from these hypothetical situations should be interpreted with caution because individuals may behave quite differently if given significant, monetary incentives.

This article argues that emotion is an important aspect of the human condition that can enhance decision making.

Shefrin and Statman (1985) argue that regret is an important factor explaining the disposition effect—the tendency to sell superior-performing stocks too early and hold on to losing stocks too long. Shefrin and Statman include Kahneman and Tversky's prospect theory as a framework to explain why investors might sell winners too early while holding on to losers. According to Shefrin and Statman, investors are more likely to realize gains than losses. The fear of regret leads investors to postpone losses whereas, symmetrically, the desire for pride leads to the realization of gains. An individual experiences regret when closing a position with a loss because of the poor investment decision. Conversely, an individual feels pride or elation when closing a position with a gain because his financial decision resulted in a profit.

Standard economic models of choice can be extended to incorporate emotions, including regret. For example, in Hermalin and Isen's (2000) model, individuals are fully rational and maximize the discounted value of future utility. Emotions directly enter utility functions, with negative emotions, such as guilt or regret, reducing utility. This research takes emotions as given, rather than trying to explain why people have emotions, and concludes that incorporating the psychological finding that emotion affects decision making into models of rational behavior gives important insight into behavioral phenomena.

Emotional Reactions

Thus far, this article has argued that emotional disposition, including regret, can affect financial decision making. Emotional responses are also induced by the plethora of stimuli people encounter every day. An individual's affective assessment is the sentiment that arises from the stimulus. For instance, when an individual is negotiating with another party and experiences a feeling of dislike for the other party, the outcome of the negotiation is likely affected. Thus, *affect* refers to the quality of a stimulus and reflects a person's impression or assessment. Cognitively, an individual's perception includes affective reactions so that judgment and decision making are inextricably linked to these reactions.

Arguably, people's thoughts are made up of images that include perceptual and symbolic representations (Damasio 1994; Charlton 2000). The images are marked by positive or negative feelings that are linked to somatic (or bodily) states. At the neural level, somatic markers arising from experience establish a connection between an entity or event and a body state (pleasant or unpleasant). In effect, affective reactions are cognitive representations of distinct body states. People are attracted to stimuli associated with positive somatic markers and steer away from those asso-

ciated with negative somatic markers. Readily available affective reactions provide expedient means for decision making because they make it far easier to weigh the pros and cons of alternative stimuli (Finucane et al. 2000).

Research that directly examines the role of affect in financial decisions is limited. More research is warranted because affective reactions influence judgment and decision making, even without cognitive evaluations (Zajonc 1980, 1984). Furthermore, when affective reactions and cognitive evaluations diverge, the emotional aspects can exert a dominating influence on behavior (Ness and Klass 1994; Rolls 1999).

In the financial realm, MacGregor et al. (2000) conclude that there is a relationship between the image of a market and what has occurred in the market. In their experiments, participants' willingness to invest in a firm was influenced by affective reactions to the firm's industry membership. Ackert and Church (2002) also examine the portfolio allocation decisions of participants in financial experiments with selective information disclosures concerning available investment alternatives. Again, affective assessments have significant effects on decision making. Other work recognizes that affect is important in understanding managers' financial decisions. Kida, Moreno, and Smith's (2001) experimental results indicate that when making capital budgeting decisions, individuals are more likely to reject projects that elicit negative emotions. Insight into market reactions awaits further investigation.

Conclusion

This article has suggested that although emotion has important influences on financial behavior, it does not contaminate judgment. Some have called for a new paradigm, one that incorporates behavioral influences and better models actual behavior. Without question, the traditional finance paradigm has been challenged. Many anomalies have been reported. Yet a paradigm is rarely displaced by anomalies (Kuhn 1970). If a paradigm is to be replaced, it must be replaced by another paradigm that provides a superior explanation of the facts. According to Kuhn, "so long as the tools a paradigm supplies continue to prove capable of solving the problems it defines, science moves fastest and penetrates most deeply through confident employment of those tools. The reason is clear. As in manufacture so in science—retooling is an extravagance to be reserved for the occasion that demands it. The significance of crises is the indication they provide that an occasion for retooling has arrived" (1970, 76). Has the time for retooling in finance reached Kuhn's crisis level?

Though recent models explain certain aspects of financial decision making that appear to be inconsistent with the efficient market hypothesis, financial economists are without a superior paradigm. Yet that is not to suggest that emotional behavior should be ignored. While some argue that in certain situations emotion may "get in the way" and lead to suboptimal decision making, we believe that emotion is an important aspect of the human condition that can actually enhance decision making.

6. Hypothetical choices may not be consistent with choices made when the incentives are real. In their experiments, Holt and Laury (2002) show that subjects are considerably more risk averse when payoffs are in cash rather than hypothetical.

REFERENCES

Ackert, Lucy F., and Bryan K. Church. 2001. The effects of subject pool and design experience on rationality in experimental asset markets. *Journal of Psychology and Financial Markets* 2, no. 1:6–28.

———. 2002. Affective evaluation and individuals' investment decisions. Kennesaw State University and Georgia Tech, unpublished working paper.

Arnold, M.B. 1960. *Emotion and personality.* New York: Columbia University Press.

Ashbury, F. Gregory, Alice M. Isen, and And U. Turken. 1999. A neuropsychological theory of positive affect and its influence on cognition. *Psychological Review* 106, no. 3:529–50.

Barberis, Nicholas, Ming Huang, and Tano Santos. 2001. Prospect theory and asset prices. *Quarterly Journal of Economics* 116, no. 1:1–53.

Black, Fischer. 1986. Noise. *Journal of Finance* 41, no. 3:529–43.

Bradley, Margaret M., and Peter J. Lang. 2000. Measuring emotion: Behavior, feeling, and physiology. In *Cognitive neuroscience of emotion*, edited by Richard D. Lane and Lynn Nadel. New York: Oxford University Press.

Charlton, Bruce G. 2000. *Psychiatry and the human condition*. Oxford: Radcliffe Medical Press.

Chen, Shu-Heng, and Chia-Hsuan Yeh. 2002. On the emergent properties of artificial stock markets: The efficient markets hypothesis and the rational expectations hypothesis. *Journal of Economic Behavior and Organization* 49, no. 2:217–39.

Damasio, Antonio R. 1994. *Descartes' error: Emotion, reason, and the human brain.* New York: Putnam.

Daniel, Kent, David Hirshleifer, and Siew Teoh. 2001. Investor psychology in capital markets: Evidence and policy implications. Northwestern University working paper.

Eisenberg, Amy E., Jonathan Baron, and Martin E.P. Seligman. 1998. Individual differences in risk aversion and anxiety. University of Pennsylvania, unpublished working paper.

Elster, Jon. 1998. Emotions and economic theory. *Journal of Economic Literature* 36, no. 1:47–74.

Fama, Eugene F. 1970. Efficient capital markets: A review of theory and empirical work. *Journal of Finance* 25, no. 2:383–417.

———. 1991. Efficient capital markets: II. *Journal of Finance* 46, no. 5:1575–1617.

Finucane, Melissa L., Ali Alhakami, Paul Slovic, and Stephen M. Johnson. 2000. The affect heuristic in judgments of risks and benefits. *Journal of Behavioral Decision Making* 13, no. 1:1–17.

Frank, Robert H. 1988. *Passions within reason*. New York: Norton.

Frijda, Nico H. 2000. The psychologists' point of view. In *Handbook of emotions*, edited by Michael Lewis and Jeannette M. Haviland-Jones, 2d ed. New York: The Guilford Press.

Goleman, Daniel. 1995. *Emotional intelligence*. New York: Bantam.

Haugen, Robert A. 1995. *The new finance: The case against efficient markets*. Englewood Cliffs, N.J.: Prentice Hall.

Hermalin, Benjamin, and Alice M. Isen. 2000. The effect of affect on economic and strategic decision making. Johnson Graduate School of Management, Cornell University working paper, July.

Hirshleifer, David. 2001. Investor psychology and asset pricing. *Journal of Finance* 56, no. 4:1533–97.

Hirshleifer, David, and Tyler Shumway. 2003. Good day sunshine: Stock returns and the weather. *Journal of Finance* 58, no. 3:1009–32.

Hockey, G. Robert J., A. John Maule, Peter J. Clough, and Larissa Bdzola. 2000. Effects of negative mood states on risk in everyday decision making. *Cognition and Emotion* 14, no. 6:823–56.

Hogarth, Robin M., and Melvin W. Reder. 1986. Editors' comments: Perspectives from economics and psychology. *Journal of Business* 59, no. 4:S185–S207.

Holt, Charles A., and Susan K. Laury. 2002. Risk aversion and incentive affects. *American Economic Review* 92, no. 5:1644–55.

Isen, Alice M. 1999. Positive affect. In *Handbook of cognition and emotion*, edited by Tim Dagleish and Mick Power. New York: John Wiley and Sons.

———. 2000. Positive affect and decision-making. In *Handbook of emotions*, edited by Michael Lewis and Jeannette M. Haviland-Jones. 2d ed. New York: The Guilford Press.

Isen, Alice M., Thomas E. Nygren, and F. Gregory Ashby. 1988. Influence of positive affect on the subjective utility of gains and losses: It is just not worth the risk. *Journal of Personality and Social Psychology* 55 (November): 710–17.

Jamal, Karim, and Shyam Sunder. 1996. Bayesian equilibrium in double auctions populated by biased heuristic traders. *Journal of Economic Behavior and Organization* 31, no. 2:273–91.

———. 2001. Why do biased heuristics approximate Bayes rule in double auctions? *Journal of Economic Behavior and Organization* 46, no. 4:431–35.

Jensen, Michael. 1978. Some anomalous evidence regarding market efficiency. *Journal of Financial Economics* 6, no. 2/3:95–101.

Kahneman, Daniel, and Amos Tversky. 1979. Prospect theory: An analysis of decision-making under risk. *Econometrica* 47, no. 2:171–85.

Kamstra, Mark, Lisa A. Kramer, and Maurice D. Levi. 2003. Winter blues: A SAD stock market cycle. *American Economic Review* 93, no. 1:324–43.

Keynes, John Maynard. 1964. *The general theory of employment, interest, and money*. New York: First Harbinger Edition, Harvest/Harcourt Brace Jovanovich.

Kida, Thomas E., Kimberly K. Moreno, and James F. Smith. 2001. The influence of affect on managers' capital-budgeting decisions. *Contemporary Accounting Research* 18, no. 3:477–94.

Kuhn, Thomas S. 1970. *The structure of scientific revolutions*. 2d ed. Chicago: University of Chicago Press.

LeDoux, Joseph. 1996. *The emotional brain: The mysterious underpinnings of emotional life*. New York: Simon & Schuster.

———. 2002. *Synaptic self: How our brains become who we are*. New York: Viking.

Lee, Charles M.C., Andrei Shleifer, and Richard H. Thaler. 1991. Investor sentiment and the closed-end fund puzzle. *Journal of Finance* 46, no. 1:75–109.

Lo, Andrew W., and Dmitry V. Repin. 2001. The psychophysiology of real-time financial risk processing. NBER Working Paper Series #8508.

MacGregor, Donald G., Paul Slovic, David Dreman, and Michael Berry. 2000. Imagery, affect, and financial judgment. *Journal of Psychology and Financial Markets* 1, no. 2:104–10.

Ness, R.M., and R. Klass. 1994. Risk perception by patients with anxiety disorders. *Journal of Nervous and Mental Disease* 182:466–70.

Raghunathan, Rajogopal, and Michel Tuan Pham. 1999. All negative moods are not equal: Motivational influences of anxiety and sadness on decision making. *Organizational Behavior and Human Decision Processes* 79, no. 1:56–77.

Rolls, Edmund T. 1999. *The brain and emotion*. Oxford: Oxford University Press.

Romer, Paul M. 2000. Thinking and feeling. *American Economic Review* 90, no. 2:439–43.

Shefrin, Hersh, and Meir Statman. 1985. The disposition to sell winners too early and ride losers too long: Theory and evidence. *Journal of Finance* 40, no. 3:777–90.

Shiller, Robert J. 2000. *Irrational exuberance*. Princeton, N.J.: Princeton University Press.

Shleifer, Andrei. 2000. *Inefficient markets: An introduction to behavioral finance*. Clarendon Lectures in Eco-nomics. Oxford: Oxford University Press.

Simon, Herbert A. 1967. Motivational and emotional controls of cognition. *Psychological Review* 74, no. 1:29–39.

Solomon, Robert C. 2000. The philosophy of emotions. In *Handbook of emotions*, edited by Michael Lewis and Jeannette M. Haviland-Jones. 2d ed. New York: The Guilford Press.

Thaler, Richard, and Eric Johnson. 1990. Gambling with the house money and trying to break even: The effects of prior outcomes on risky choice. *Management Science* 36, no. 6:643–60.

Wright, William F., and Gordon H. Bower. 1992. Mood effects on subjective probability assessment. *Organizational Behavior and Human Decision Processes* 52:276–91.

Zajonc, R.B. 1980. Feeling and thinking: Preferences need no inferences. *American Psychologist* 35, no. 2:151–75.

———. 1984. On the primacy of affect. *American Psychologist* 39 (February): 117–23.

Inflation Persistence: How Much Can We Explain?

PAU RABANAL AND JUAN F. RUBIO-RAMÍREZ

Rabanal is an economist in the monetary and financial systems department at the International Monetary Fund in Washington, D.C. Rubio-Ramírez is an economist in the research department at the Federal Reserve Bank of Atlanta. *The authors thank Tom Cunningham, Karsten Jeske, and Ellis Tallman for useful comments.*

Monetary policy is a controversial topic. Economists are still divided into two factions: those who believe that monetary policy does have real (inflation-adjusted) effects and those who are convinced that it affects only nominal variables, that is, nominal interest rates and prices. Until recently, almost any macroeconomic model in which monetary policy has real effects was based on the assumption that expectations are formed in an adaptative way, implying that agents do not use all available information when making a decision. Critics of these models argue that, given this assumption, agents are not rational and as a result allow the monetary authority to trick them over and over.

In response to this important critique, a whole class of models—New Keynesian models—has been recently proposed. These types of models combine "old" Keynesian elements (imperfect competition and short-term nominal rigidities) with a dynamic general equilibrium environment (where prices and quantities are such that markets clear) in which agents form their expectations rationally.[1] The idea behind this approach is that when short-term prices are "sticky" or "rigid"—that is, when they adjust only slowly to market shortages or surpluses—a decrease in the nominal interest rate also implies a decrease in the real interest rate. Therefore, the consumption and investment components of aggregate demand increase, implying an increase in output. But over time the excess aggregate demand shifts prices upward, thereby restoring the level of output to its potential. A drawback of the simplest version of such models (in which only one type of nominal rigidity, either sticky prices or wages, is considered) is that it does not seem to be able to reproduce the observed persistence of inflation.

The objective of this article is to determine whether adding sticky wages to a basic sticky-price model overcomes this drawback. The analysis shows that this addition "partially" solves the problem. Empirical work at the micro level suggests that the average duration of price and wage contracts is typically three to six quarters. Chari, Kehoe, and McGrattan (1998) find that, in order to match the persistence of output changes to a monetary shock, their model must assume an implausible degree (ten quarters) of price stickiness, even when capital accumulation and adjustment costs of capital are introduced. Fuhrer and Moore (1995) also show that, in a model using a reasonable length of wage contracts, it is not possible to obtain the inflation persistence observed in the data.

As Galí and Gertler (1999) point out, these models imply an aggregate supply relationship (the new Phillips curve) that relates current inflation with expectations of future inflation and real unit-labor costs. Hence, the persistence of price inflation in New Keynesian models is driven by the persistence

of real unit-labor costs. The problem of models with only one type of nominal rigidity is that, even with long-duration (sixteen quarters) price or wage contracts, real wages are still flexible and cannot induce enough persistence of inflation.

How can the baseline sticky-price model be modified so that the induced persistence of inflation in response to a monetary shock increases under plausible degrees of price or wage stickiness? A straightforward path would be to introduce some kind of backward-looking behavior in the determination of inflation. However, introducing backward-looking behavior implies departing from the assumption of rational expectations, and the Lucas (1976) critique applies.[2]

Economists are still divided into two factions: those who believe that monetary policy does have real (inflation-adjusted) effects and those who are convinced that it affects only nominal variables.

This article takes an alternative approach, exploring whether the combination of staggered price and wage setting, as in Erceg, Henderson, and Levin (2000), can match the inflation persistence observed in the data if reasonable durations of price and wage contracts are assumed. In this case, the forward-looking nature of the model is preserved. When both prices and nominal wages are sticky, so is the real wage, and therefore inflation persistence should increase.

This article analyzes whether adding staggered wage settings to the baseline sticky-price model solves the persistence-of-inflation problem when plausible durations of price and wage contracts are assumed. The analysis will show that, for a given duration of price contracts, real wage persistence significantly increases with the duration of wage contracts. This exercise is equivalent to adding sticky prices to a model with only staggered nominal wages. The exercise presented here is chosen because models containing only sticky prices are more widely used in the literature than those containing only staggered nominal wages.

Both the baseline sticky-price and the sticky-price and sticky-wage models have three main equations: an aggregate supply relationship (the new Phillips curve), an IS type of equation, and a monetary policy rule.[3] As mentioned above, the new Phillips curve relates current inflation with expectations of future inflation and real unit-labor costs. Understanding the inflation–real wage link is important in understanding why adding staggered wages to the baseline sticky-price model may solve the lack of persistence of inflation in these models. The IS curve relates output and the real rate of interest negatively, as in the undergraduate textbook version of the Keynesian model, and the IS curve includes expectations of future output. These two relationships manifest the forward-looking nature of the New Keynesian models, in which expectations are rational. To complete the model, a monetary policy rule is needed. Typically, it is modeled as an interest rate rule in which the short-term interest rate reacts to inflation and output gaps;[4] the nominal amount of money is determined from the money demand equation. Following the literature, this article uses an interest rate rule that relates today's nominal interest rates to past nominal interest rates through an interest rate–smoothing parameter. One might interpret this parameter as reflecting monetary policymakers' perceived aversion to moving the nominal rate by large steps.

The analysis in this article reveals the following: First, as most of the literature has proved, when only sticky prices and plausible-duration price contracts are considered, the model is not able to replicate the inflation persistence observed in the data. Second, in the baseline sticky-price model most of the persistence is driven by the exogenous nominal interest rate–smoothing parameter. Finally, when sticky wages are added to the baseline sticky-price model, it is possible for the inflation data autocorrelations to be reasonable approximations closely matched by the model.

The first part of the article analyzes the equations that describe a general equilibrium model with sticky prices. The discussion then shows how these equations are modified when staggered wages are added to the baseline sticky-price model. Next, the analysis examines how different parameterizations affect price-inflation persistence in the baseline sticky-price model. Finally, the study considers how those conclusions are affected when both sticky prices and wages are considered.

The Model

The baseline sticky-price model presented in this section merges Keynesian assumptions, such as imperfect competition and nominal rigidities, with the methodological advances in modern macroeconomic theory. As in traditional Keynesian models, mone-

tary policy affects real variables in the short run. Unlike the traditional Keynesian models, in New Keynesian models the equations come from an optimization process of rational agents. Two models are considered: first, a model with sticky prices but flexible wages and then a model that introduces staggered wage setting into this baseline environment.

The baseline sticky-price model. Following Blanchard and Kiyotaki (1987), the model consists of

- a large number of identical households each supplying labor services,
- a large number of intermediate-good producers producing a specific good that is an imperfect substitute for the other goods, and
- a large number of identical, competitive final-good producers.

Households consume the final good, intermediate-good producers use labor services in their production process, and final-good producers use the intermediate good in their production of the final good. The model also assumes imperfect competition in the intermediate-good markets. Thus, each intermediate-good producer chooses its price, taking as given all other good prices and wages. The intermediate-good production sector suffers an aggregate technology shock that is common across firms. For this sector, the model assumes a linear production function in labor such that the marginal product of labor is equal to the technology shock.

On the monetary policy side, the model assumes that the central bank sets the nominal interest rate through a Taylor rule and supplies as much money as households demand. The Taylor rule relates today's nominal interest rate to past nominal interest rates, inflation, and output gaps. The model also assumes that monetary policy, that is, the Taylor rule, suffers from a monetary perturbation. This perturbation reflects the difference between the information that the monetary authority has when making decisions on interest rates and the information that the researcher can observe.

Intermediate-good producers face a Calvo-type restriction when setting prices: In any given period of time, each intermediate-good producer receives a signal that allows her to change the price. This signal arrives with probability $1 - \theta_p$ and thus with probability θ_p that she must keep last period's price. The reason the Calvo-type assumption has become so popular is its simplicity. Because the probability of receiving the "green light" signal is independent of the past history of signals, the pricing decisions of firms are identical. Therefore, one does not need to keep track of each firm's pricing decision to know the aggregate price outcome, and aggregation is simple.

The intuition behind this idea is as follows: Firms face some type of "menu cost" when they want to change prices, so they cannot change prices every period. In this environment the probability that a firm has its price fixed for one period is $1 - \theta_p$, for two periods is $\theta_p (1 - \theta_p)$, for three is $\theta_p^2(1 - \theta_p)$, and so on. Given these probabilities, the average number of periods that prices are going to be fixed can be calculated. Hence, this average duration of a price contract is equal to $[1 - \theta_p] + 2[\theta_p(1 - \theta_p)] + 3[\theta_p^2(1 - \theta_p)] + ... = 1/(1 - \theta_p)$. It is important to remember the relationship between θ_p and the average duration of price contracts.

This analysis will not go through the derivation of the main equations. (The reader is referred to Rabanal and Rubio-Ramírez 2003.) Instead, the discussion will introduce the key relationships and give some intuition. In all cases, the variables are expressed in logarithmic terms. Let y_t denote output; w_t, the nominal wage; p_t, the price level; and Δp_t, the price inflation rate.

The model is represented by the following set of equations:

$$(1)\quad y_t = -\frac{1}{\sigma}(r_t - E_t\Delta p_{t+1} - \rho_\beta) + E_t y_{t+1};$$

$$(2)\quad \Delta p_t = \beta E_t \Delta p_t + \kappa_p(w_t - p_t - a_t + \mu);$$

$$(3)\quad w_t - p_t = \vartheta + mrs_t;$$

$$(4)\quad mrs_t = (\sigma + \gamma)y_t - \gamma a_t,$$

and

$$\kappa_p = \frac{[(1-\theta_p\beta)(1-\theta_p)]}{\theta_p};$$

1. For another way of answering this critique, see Christiano and Eichenbaum (1992).
2. The Lucas critique implies that any Federal Reserve policy change will affect consumers' expectations, so the Federal Reserve cannot take consumers' expectations as constant.
3. An IS equation relates output today with output tomorrow as a function of the nominal interest rate and inflation.
4. Even though this article does not do so, it is also possible to model the interest rate rule as forward looking in the sense that it reacts to expected future inflation and output gaps. However, simulations suggest that our results would remain basically unchanged. Output gaps are the difference between actual and potential output.

(5) $a_t = \rho_a a_{t-1} + \varepsilon_t^a$,

where β is the discount factor, ρ_β is equal to $\log(\beta)$, γ is the inverse of the elasticity of labor supply to real wage, σ is the inverse of the intertemporal elasticity of substitution, r_t is the nominal interest rate, mrs_t is the marginal rate of substitution between consumption and worked hours, μ is the desired markup on marginal product, w_t is the hourly wage, ϑ is the desired markup on the real wage, κ_p is the elasticity of inflation to the marginal cost, and a_t is the aggregate productivity shock. It is assumed that $\varepsilon_t^a \sim \text{iid}N(0, \sigma_a)$.

Equation (1) is a log-linearized version of the Euler equation, which arises from the household's optimal saving-consumption decision, after imposing the clearing market condition that consumption equals output. From equation (1) it is clear that the higher the nominal interest rate, r_t, or the lower tomorrow's expected inflation, $E_t\Delta p_{t+1}$, the lower today's output, y_t, and the higher the savings.

The baseline sticky-price model merges Keynesian assumptions, such as imperfect competition and nominal rigidities, with the methodological advances in modern macroeconomic theory.

Equation (1) can also be iterated forward to yield

(6) $$y_t = -\frac{1}{\sigma} E_t \sum_{\tau=0}^{\infty} (r_{t+\tau} - \Delta p_{t+1+\tau} - \rho_\beta).$$

This iteration shows that output depends on current and expected future gaps between the real interest rate and its long-run value. Thus, one concludes that output is at its steady-state value only when real interest rates differ by the log of the discount factor and are expected to do so. In other words, when the real interest rate is high, savings are high and consumption (and, hence, output) is low; when the real interest rate is low, savings are low and consumption is high.

Equation (2) is called the New Keynesian Phillips curve, and it is obtained from the aggregation of price-optimal decisions of firms. Price inflation depends on tomorrow's expected price inflation, $E_t\Delta p_{t+1}$, and the percentage deviation of real wage, $w_t - p_t$, from the desired markup over the marginal product of labor, $a_t - \mu$, where μ is the desired markup on the marginal product and depends on the elasticity of substitution between different types of intermediate goods used to produce the final good. This equation is the most important piece of the New Keynesian models. As mentioned above, until the introduction of these models, almost any setup able to generate short-term real effects of monetary policy was based on backward-looking behavior. As equation (2) shows, this situation is no longer true: In this environment, inflation has a forward-looking root, and monetary policy affects output through its effects on future real interest rates and real wages.

If equation (2) is solved forward, the resulting equation is

(7) $$\Delta p_t = -\kappa_p E_t \sum_{\tau=0}^{\infty} (w_{t+\tau} - p_{t+\tau} - a_{t+\tau} + \mu).$$

It reveals that price inflation depends on current and expected future gaps between real wages and the desired markup over the marginal product of labor. Thus, one concludes that price inflation is at its steady-state value only when real wages and the marginal product of labor differ by the desired markup and are expected to do so. If firms do not expect wages to increase over the marginal product of labor more than the desired markup, they will not increase prices, and inflation will be at its steady-state value.

Equation (3) relates the real wage, $w_t - p_t$, the marginal rate of substitution between consumption and worked hours, mrs_t, and the desired real-wage markup, ϑ, that depends on the elasticity of substitution between different types of labor used to produce each intermediate good. Equation (4) relates the marginal rate of substitution between consumption and worked hours, mrs_t, with output, y_t, and the aggregate productivity shock, a_t. This expression is obtained by imposing the clearing market condition that consumption equals total production and by using the production function that relates hours worked with output and the productivity shock. Equation (5) shows how the aggregate productivity shock, a_t, (or technology shock) evolves over time.

A monetary policy rule is needed to complete the general equilibrium model. This analysis will consider a Taylor-type rule with the following formulation:

(8) $r_t = \rho r_{t-1} + (1-\rho)(\gamma_\pi \Delta p_t + \gamma_x y_t) + \varepsilon_t$,

where γ_π and γ_x are the elasticities of the nominal interest rate to current price inflation and output gap. ε_t is the monetary shock, and it is independent and identically distributed normally with zero mean and standard deviation σ_r.

Equation (8) relates today's nominal interest rate, r_t, to yesterday's nominal interest rate, r_{t-1}, price inflation, Δp_t, and output gap, y_t. It is assumed that ρ is between 0 and 1, $\gamma_\pi > 1$, and $\gamma_x > 0$. The interest rate–smoothing coefficient is included in the Taylor rule mainly for empirical reasons (see the paper by Clarida, Galí, and Gertler 2000). In addition, Woodford (2002) provides some theoretical background about why the central bank might be interested in smoothing interest rates. In this way, the nominal interest rate will have some exogenously driven persistence. This model imposes the condition $\gamma_\pi > 1$: The monetary authority increases the nominal interest rate more than one to one with respect to inflation to induce a unique, stationary solution to the system (see Woodford 2002).

In the baseline sticky-price model, and in the sticky-price and -wage model presented in the next section, two sources of uncertainty exist: one is technological, ε_t^a, and the other is monetary, ε_t.

Two key parameters drive inflation persistence in the baseline sticky-price model: First, θ_p modifies the slope in equation (2). Hence, the larger θ_p, the longer the duration of price contracts and the higher the generated persistence of inflation. The second, ρ, is the interest rate–smoothing coefficient. A higher ρ increases the persistence of both monetary shocks and output gaps, also making inflation more persistent. Given the importance of these two parameters, the next section demonstrates how different calibration choices for them modify the persistence of inflation that this model can generate.

The sticky-price and -wage model. As the next section of the article will show, with only sticky prices it is not enough to replicate the persistence of inflation that is observed in the data. Therefore, this section presents a version of the model with staggered prices and wages, as in Erceg, Henderson, and Levin (2000). The inclusion of nominal wage rigidities will increase the real wage and, one hopes, inflation inertia. The model setup is similar to the one presented in the last subsection. As before, the model consists of a continuum of households each supplying a specific labor service that is an imperfect substitute for the other labor services, a continuum of intermediate-good producers producing a specific good that is an imperfect substitute for the other goods, and a continuum of identical competitive final-good producers. As in the baseline sticky-price model, households consume the final good, intermediate-good producers use labor services in their production process, and final-good producers use the intermediate good in their production of the final good. The model also assumes imperfect competition on the intermediate-good markets. Thus, each intermediate-good producer chooses its price, taking as given all other good prices and wages. The intermediate-good production sector suffers an aggregate technology shock that is common across firms. For this sector, a linear production function in labor is assumed, so the marginal product of labor is equal to the technology shock. Finally, the central bank also sets the nominal interest rate, through a Taylor rule, and supplies as much money as households demand. Also, as in the previous model, it is assumed that the Taylor rule suffers from a monetary perturbation.

Just as in the baseline sticky-price model, producers of intermediate goods face a Calvo-type restriction when setting prices, as described earlier. In this new model, households face an additional Calvo-type restriction when setting their wages. In this environment the probability that a household has its wage fixed for one period is $1 - \theta_w$. Therefore, the average duration of a wage contract is equal to $1/(1 - \theta_w)$.

With only sticky prices it is not enough to replicate the persistence of inflation that is observed in the data.

The model can be represented by the following set of equations:

$$(9)\quad y_t = -\frac{1}{\sigma}(r_t - E_t\Delta p_{t+1} - \rho_\beta) + E_t y_{t+1};$$

$$\Delta p_t = \beta E_t \Delta p_{t+1} + \kappa_p(w_t - p_t - a_t + \mu);$$

$$(10)\quad \Delta w_t = \beta E_t \Delta w_{t+1} + \kappa_w(mrs_t - (w_t - p_t) + \vartheta);$$

$$w_t - p_t = w_{t-1} - p_{t-1} + \Delta w_t - \Delta p_t;$$

$$mrs_t = (\sigma + \gamma)y_t - \gamma a_t;$$

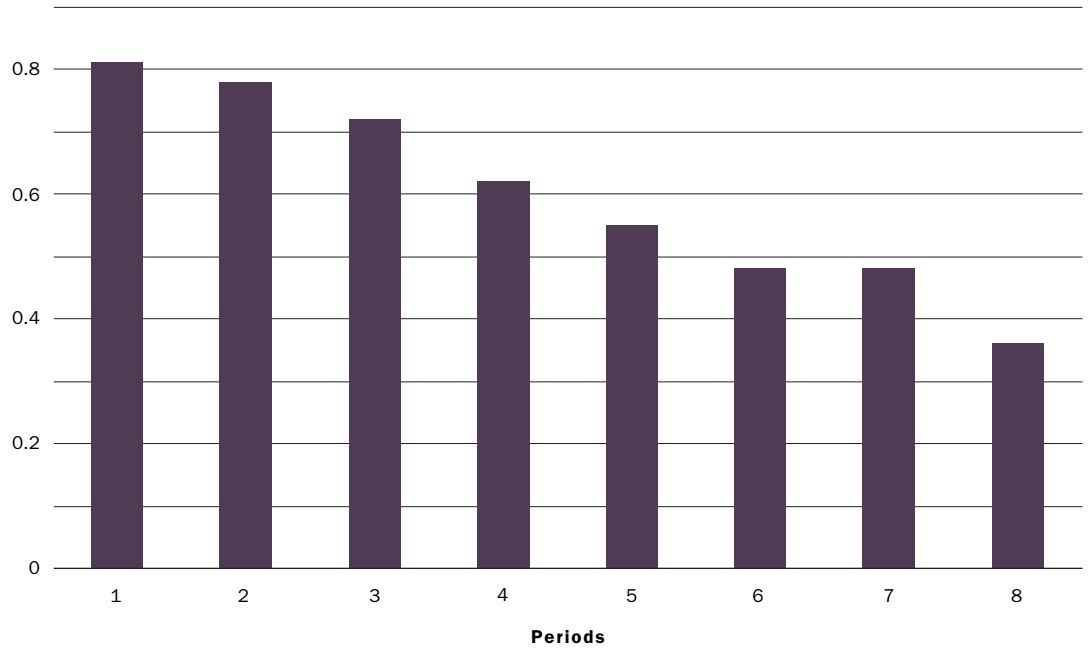
where

$$\kappa_p = \frac{[(1 - \theta_p\beta)(1 - \theta_p)]}{\theta_p};$$

$$\kappa_w = \frac{[(1 - \theta_w\beta)(1 - \theta_w)]}{\theta_w(1 + \gamma\varphi)}.$$

FIGURE 1

The Autocorrelation Function of the GDP Deflator for the Nonfarm Business Sector between 1960:01 and 2001:04



Again, a monetary policy rule is needed to complete the model. As before, a Taylor-type rule with the following structure is considered:

$$r_t = \rho r_{t-1} + (1 - \rho)(\gamma_\pi \Delta p_t + \gamma_x y_t) + \varepsilon_t.$$

If the equations that describe the baseline sticky-price model are carefully compared with those that describe the sticky-price and sticky-wage model, two differences should be apparent. First, the inclusion of sticky wages does not modify the structure of the New Keynesian Phillips curve (equations [2] and [9] are identical). Second, mrs_t is no longer equal to a markup over real wages. Instead of equation (3), equation (10) now relates wage inflation to expected wage inflation and the percentage deviation of real wages, mrs_t, from the desired markup over the real wage of labor, $(w_t - p_t) - \vartheta$, in the same way the New Keynesian Phillips curve does.

Comparison of the two models. Although in the baseline sticky-price model θ_p and ρ drive inflation, in the sticky-price and -wage model a bigger θ_w implies a longer duration of wage contracts and, hence, a more persistent real wage. The New Keynesian Phillips curve (equation [9]) implied by this new version of the model relates price inflation persistence to real wage persistence; hence, a larger θ_w implies higher inflation persistence.

In the next section, the analysis explores how different calibration choices for these three parameters modify the persistence of inflation that this model can generate. Notably, because the New Keynesian Phillips curve remains unaltered, then in either model persistence in price inflation after a monetary shock hits the economy is driven by κ_p and the persistence of the real wage, $w_t - p_t$. The inclusion of nominal wage rigidities does not modify κ_p. Hence, the addition of nominal wage rigidities only increases the price-inflation persistence if it increases the persistence of real wages.

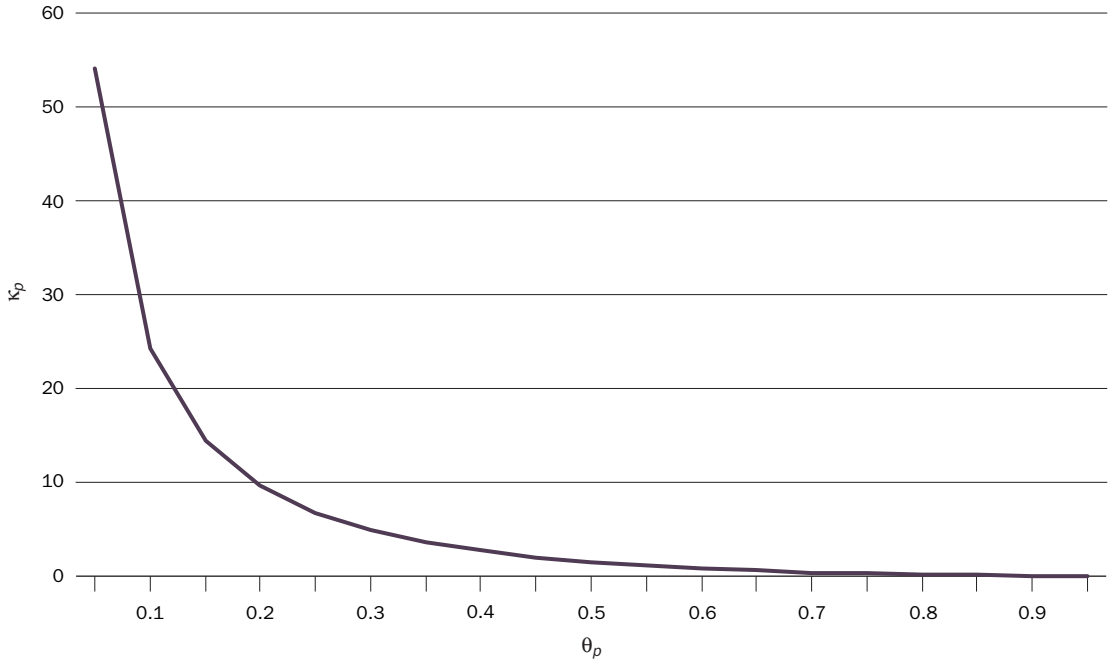
Inflation Persistence Analysis

To study the persistence of price inflation implied by the two models, this analysis first reports the observed autocorrelations of price inflation and then performs some numerical exercises to study the autocorrelation functions of price inflation implied by the basic sticky-price model when only a monetary shock is considered. Finally, the analysis does the same for the sticky-price and sticky-wage model.

To understand how much persistence in price inflation these models can generate given a plausible degree of price and wage stickiness, one could modify the parameters of the model (in particular, θ_p and θ_w) until κ_p or real-wage stickiness is such that price inflation matches observed inflation. As

FIGURE 2

The Elasticity of Inflation to Marginal Cost (κ_p) As a Function of the Probability of Price Change (θ_p) in the Sticky-Price Model



mentioned, the problem with this approach is that access to additional empirical evidence, such as surveys or data panels, provides us reasonable bounds for most of the parameters of the model. Thus, a plausible degree of price and wage stickiness means that the wage and price contract length implied by θ_p and θ_w are inside these bounds.

The autocorrelation function of price inflation. Figure 1 shows the autocorrelation function of the gross domestic product (GDP) deflator for the nonfarm business sector between 1960:01 and 2001:04. First, the autocorrelation function implied by the GDP deflator reported here is similar to the one implied by either the consumer price index (CPI) or the personal consumption expenditures index (PCE).

Second, even after five periods the autocorrelation is 0.5. The following analysis shows that New Keynesian models with only sticky prices have a number of problems replicating this slow decay of the autocorrelogram.

Persistence in the sticky-price model. The effects of θ_p on price inflation are twofold. First, it affects the slope of the New Keynesian Phillips curve:

$$\kappa_p = \frac{[(1-\theta_p\beta)(1-\theta_p)]}{\theta_p};$$

Second, θ_p affects the persistence of the percentage deviation of the real wage $(w_t - p_t)$ with respect to the marginal product of labor (a_t).

Before studying the relationship between θ_p and real wage persistence, the analysis will first concentrate on understanding how the price contract duration, $1/(1-\theta_p)$, affects κ_p. Notice the following relationship:

$$\frac{\partial \kappa_p}{\partial \theta_p} = -\frac{\sigma+\gamma}{\theta_p^2}(1-\theta_p^2\beta) < 0.$$

This derivative implies that the higher θ_p (that is, the higher the price contract duration), the lower κ_p. One can observe this relationship in Figure 2, which plots κ_p as a function of θ_p. Under the limitation $\theta_p \to 1$, that is, when prices are fixed forever, $\kappa_p \to 0$ and $\pi_t = 0$ forever, implying the highest persistence possible.

As mentioned before, the issue is that, as Figure 3 shows, the higher θ_p, the higher the average duration of price contracts, $1/(1-\theta_p)$. As many authors have reported (see, for example, Dutta, Berger, and Levy 1997; Blinder et al. 1998), observed average price change is not much longer than one year. This observation implies that analysis should be restricted to values of θ_p that imply durations no longer than five quarters, that is, $\theta_p \leq 4/5$.

FIGURE 3

Duration As a Function of the Probability of Price Change (θ_p) in the Sticky-Price Model

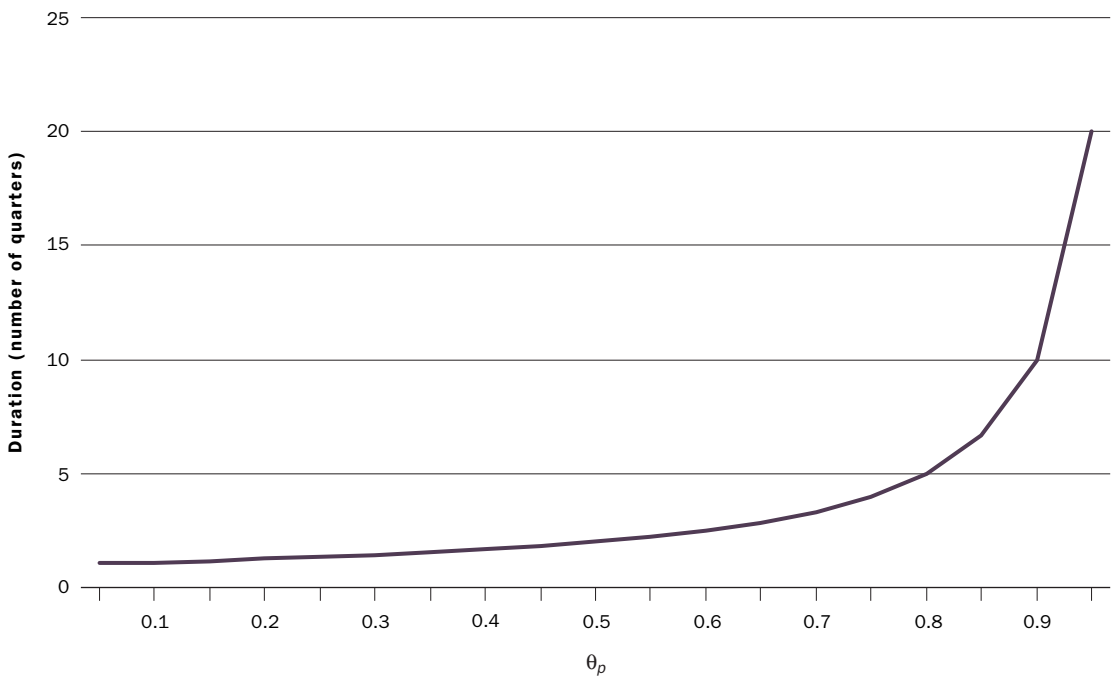


TABLE 1

Calibration for the Sticky-Price Model

Variable	Value
σ	1
γ	1
β	0.99
μ	1.2
ϑ	1.2
ρ	0.8
ρ_a	0.8
σ_r	1
σ_a	1
γ_π	1.5
γ_x	0.5
θ_p	$^3/_4$

The analysis has shown that increasing persistence by letting $\theta_p \to 1$ ($\kappa_p \to 0$) is not consistent with the evidence on price contract duration. The following numerical simulations study the effects of different parameter values of θ_p (different price contract durations) on price inflation persistence, examining the real-wage and price-inflation autocorrelation functions that the baseline sticky-price model generates under these conditions.

The baseline calibration used in the following analysis is shown in Table 1. The inverse intertemporal rate of substitution and the elasticity of labor supply, σ and γ, are set to 1. Because quarterly data are used, β being set to 0.99 implies a 4.1 percent annualized real interest rate. The calibration of μ and ϑ at 1.2 implies a 20 percent markup over marginal costs and real wages, respectively. Both ρ and ρ_a are set to 0.8, and σ_r and σ_a are set to 1. Taylor's rule elasticities, γ_π and γ_x, are set to Taylor's original guesses. θ_p is set to $^3/_4$, which implies an average duration of price contracts of four periods.

Figures 4 and 5 illustrate the autocorrelation functions of price inflation and real wages when only a monetary shock is considered for different average durations of price contracts. The longer the duration, the higher price inflation persistence. The intuition for this result is that when prices are sticky, a positive (negative) monetary shock will increase (decrease) demand and real output. The longer the average price contract lasts, the more persistent is the effect on output. As equations (3) and (4) show, real wages are linked to output, so the higher the output persistence, the higher the real wage persistence. Because no technology shock is involved, the marginal product of labor is constant, and the real wage and its deviation from the marginal product of labor exhibit the same auto-

FIGURE 4

The Autocorrelation Function of Price Inflation When Only a Monetary Shock Is Considered in the Sticky-Price Model

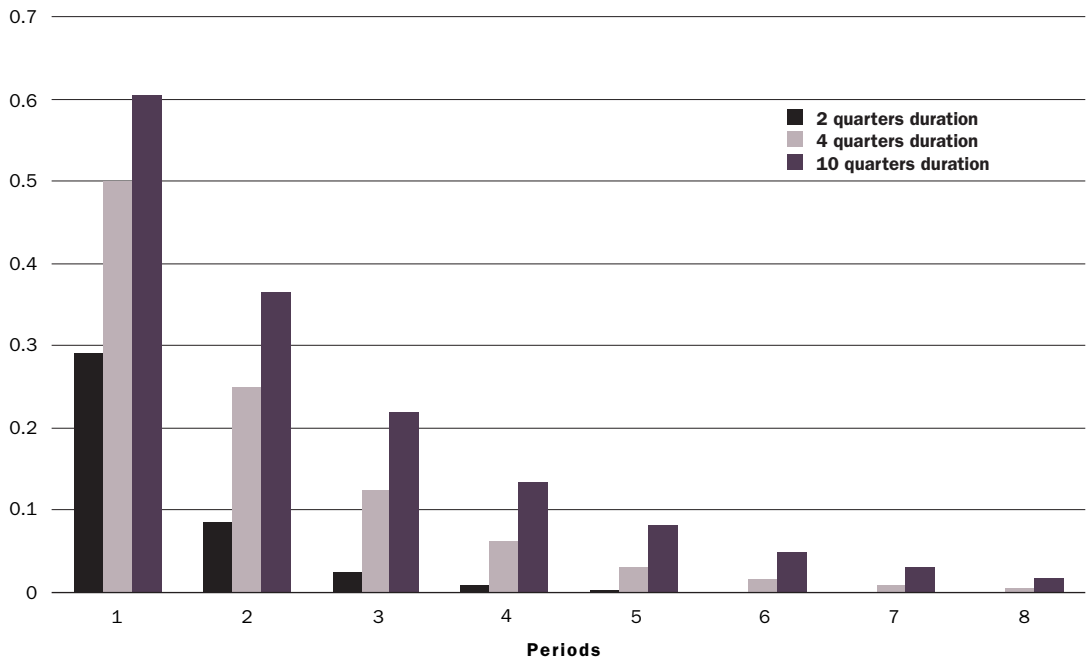


FIGURE 5

The Autocorrelation Function of Real Wages When Only a Monetary Shock Is Considered in the Sticky-Price Model

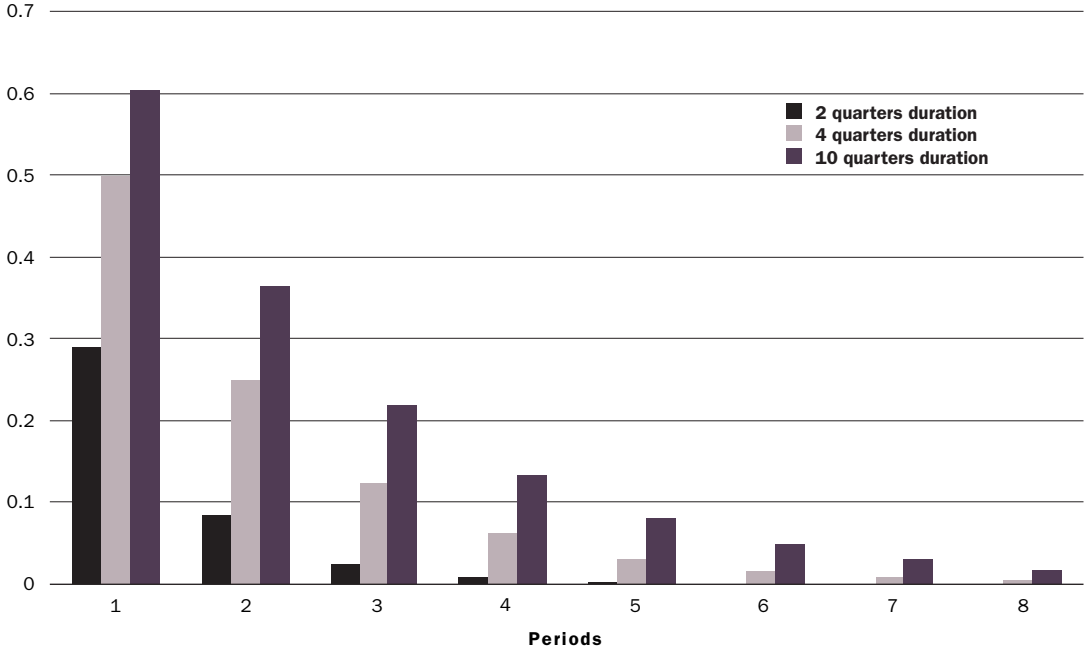


FIGURE 6

The Autocorrelation Function of Price Inflation When Only a Monetary Shock Is Considered for Different Values of ρ in the Sticky-Price Model

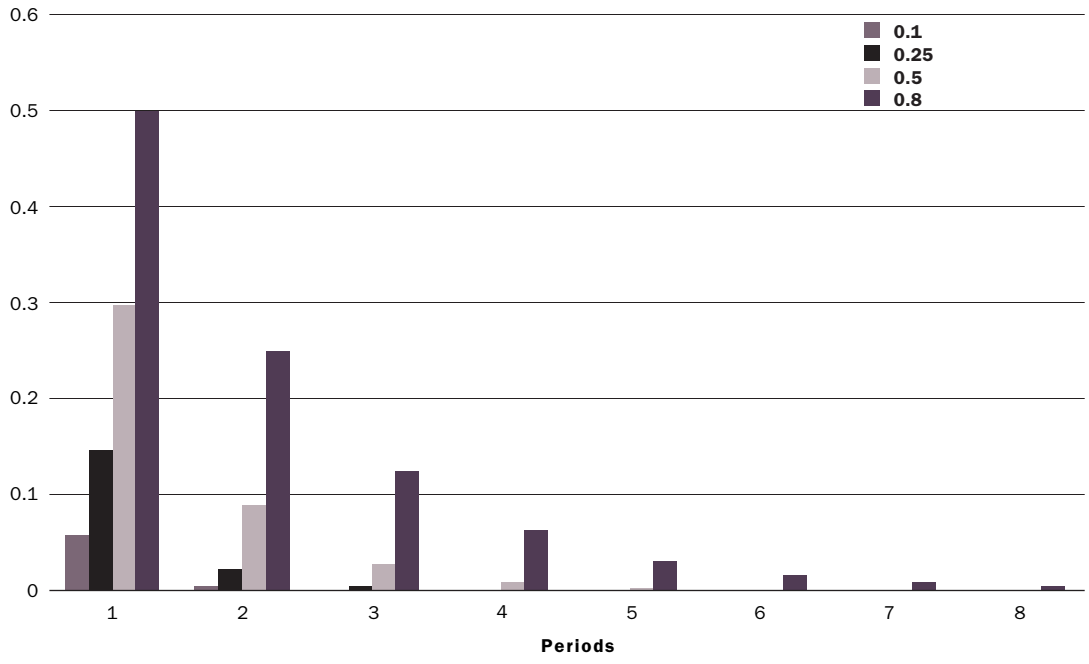


TABLE 2

Calibration for the Sticky-Price and Sticky-Wage Model

Variable	Value
σ	1
γ	1
β	0.99
μ	1.2
ϑ	1.2
ρ	0.8
ρ_a	0.8
σ_r	1
σ_a	1
γ_π	1.5
γ_x	0.5
θ_p	$^3/_4$
θ_w	$^4/_5$

correlation function. Thus, one can conclude that the longer the average contract, the more persistent is the effect on real wages and price inflation.

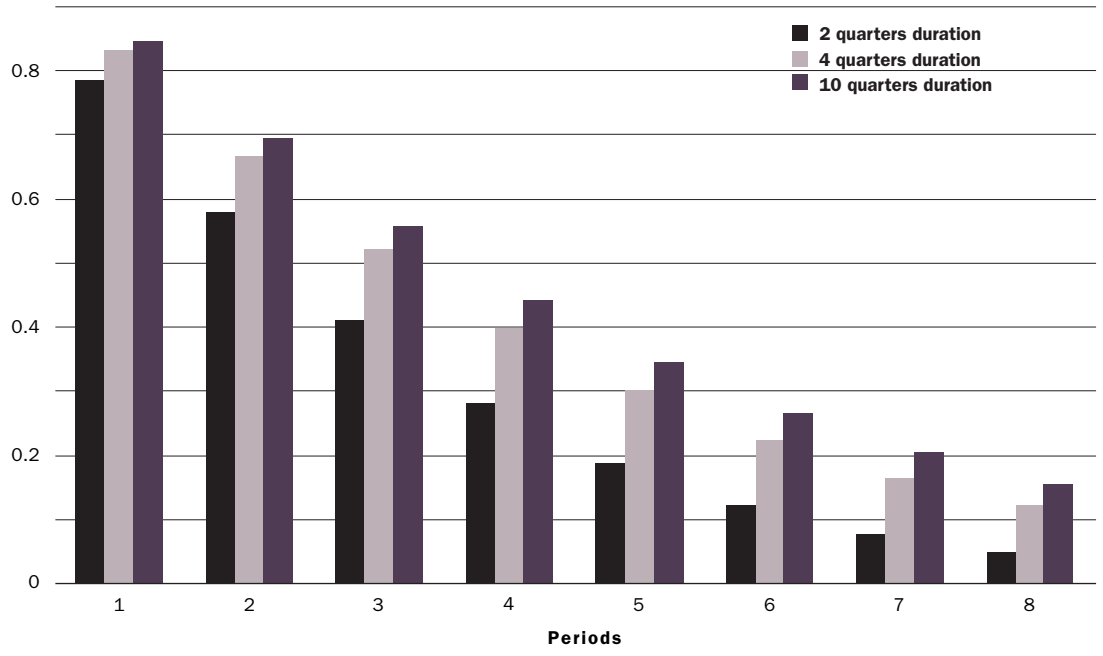
It is important to understand what the source of this persistence is. The following analysis considers how the source of exogenous nominal interest rate persistence, ρ, affects the persistence of inflation that this simple model is able to generate.

Figure 6 shows the autocorrelation function of price inflation for different values of ρ (the exogenous persistence parameter of the nominal interest rate) when only a monetary shock is considered. For low values of ρ, inflation persistence is very low. The intuition is as follows. Nominal interest rate persistence depends on ρ. As equation (6) shows, the higher (lower) the nominal interest rate persistence, the higher (lower) the output persistence. Because only a monetary shock is considered, real wages and output share the same autocorrelation function, so the higher (lower) the nominal interest rate persistence, the higher (lower) the real wage and price inflation persistence. Under these conditions, inflation persistence greatly depends on ρ, the nominal interest rate exogenous persistence. Indeed, Figure 6 shows that a model with only sticky prices does not amplify the inflation persistence of a monetary shock beyond that induced by ρ.

From this analysis one can conclude that the model with only sticky prices is not able to generate endogenous persistence beyond that obtained through the coefficient ρ. This conclusion raises the following questions: Is it possible to generate inflation persistence in this model? Is inflation persistence highly linked to ρ? Is there an obvious mechanism generating it? From the observations in Figure 6, it seems that the correct answer is that the inflation

FIGURE 7

The Autocorrelation Function of Price Inflation When Only a Monetary Shock Is Considered in the Sticky-Price and Sticky-Wage Model



persistence is highly related to ρ and that there is no other mechanism that can generate it.

These results indicate that the baseline sticky-price model is not able to generate enough endogenous inflation persistence, so the analysis next considers whether the sticky-price and sticky-wage model can do it.

Persistence in the sticky-price and sticky-wage model. As mentioned earlier, the inclusion of wage rigidities does not modify κ_p. Thus, the impact of wage rigidities on price inflation persistence should come through their effect on the persistence of the real wage. Table 2 lists the basic calibration used in this model. The parameters are set to the same values used in the baseline sticky-price model. θ_w is set to $^4/_5$, implying an average wage contract duration of five quarters.

Figures 7 and 8 show the autocorrelation function of price inflation and real wage for a given average duration of the price contract of four quarters ($\theta_p = {}^3/_4$) when different values of θ_w and just a monetary shock are considered. The inclusion of wage rigidities increases the persistence of both real wages and price inflation. In the sticky-price model, nominal wages move freely, making real wages not persistent. When both wages and prices are sticky, real wages display more persistence. As noted before, the persistence of the real wage deviation from the marginal product of labor drives price inflation persistence. In Figures 7 and 8, which consider only a money shock, the marginal product of labor does not move, and price inflation persistence also increases. One can conclude that the addition of nominal wage stickiness makes the reaction of price inflation to money shocks more persistent.

Figure 9 reports the autocorrelation function of price inflation for different values of ρ when only a monetary shock is considered. In the basic sticky-price model, the price inflation persistence depends greatly on ρ. Figure 9 shows that when both prices and wages are sticky, the persistence of inflation the model generates as a response to a monetary shock does not depend on ρ. The addition of sticky wages to the baseline sticky price model increases real wage persistence in such a way that, even with very low ρ, the model is able to generate a persistent inflation response. In addition, the introduction of staggered wage contracts to the sticky price model in a pure forward-looking model helps increase inflation persistence.

Conclusion

This article analyzes the ability of a model with both sticky prices and wages to solve one of the most important shortcomings of the baseline sticky-price model: the lack of persistence of inflation when

FIGURE 8

The Autocorrelation Function of Real Wages When Only a Monetary Shock Is Considered in the Sticky-Price and Sticky-Wage Model

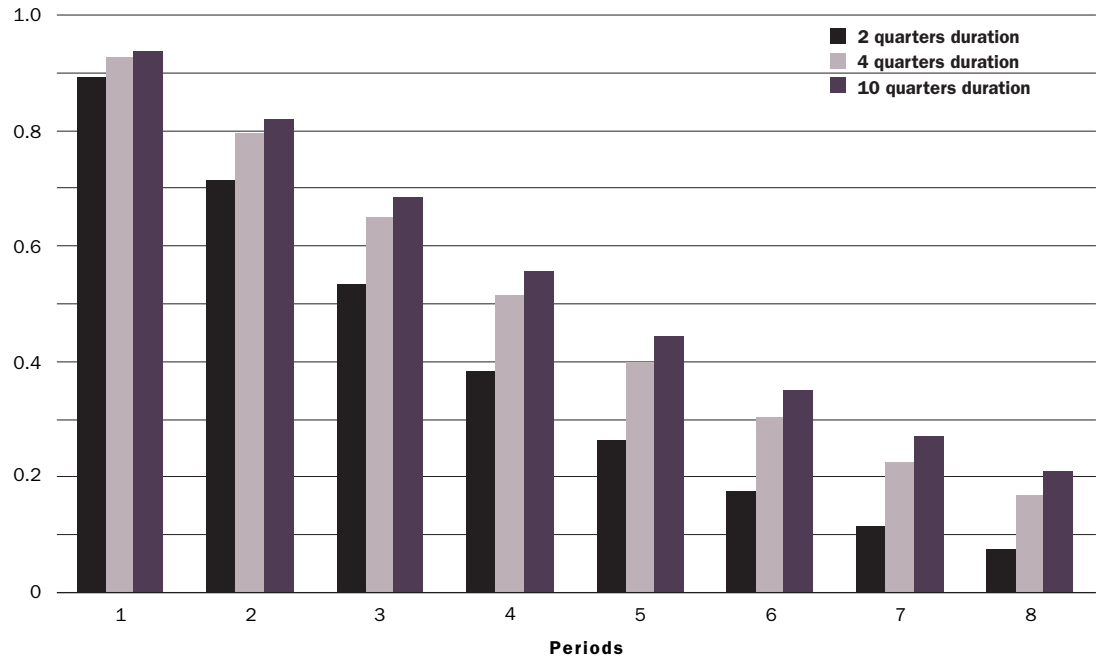
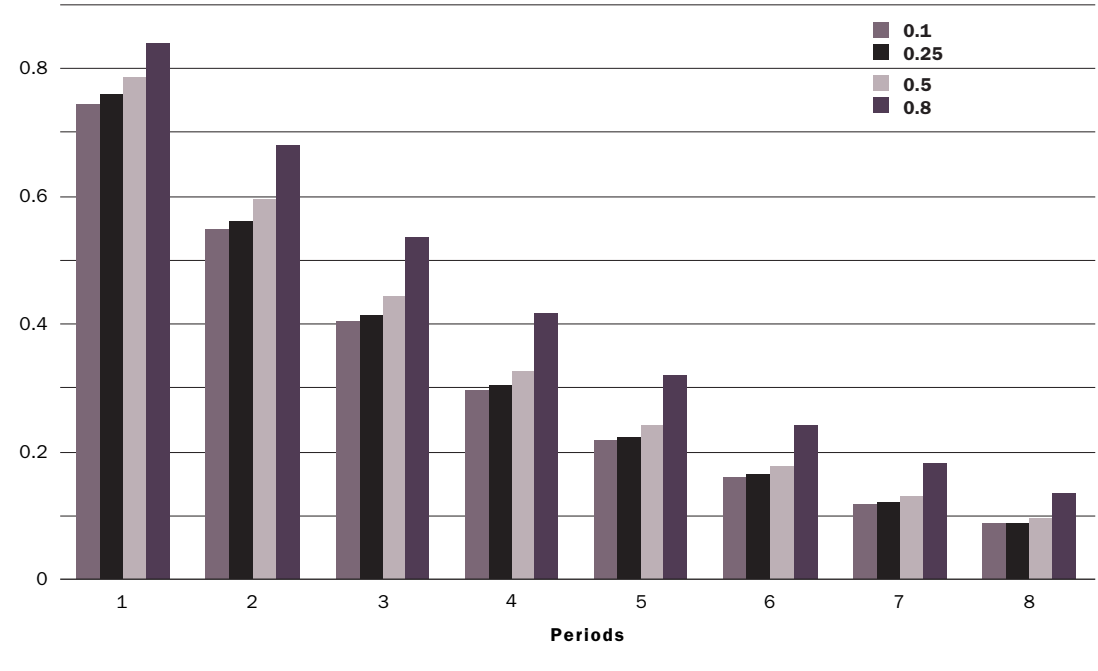


FIGURE 9

The Autocorrelation Function of Price Inflation When Only a Monetary Shock Is Considered for Different Values of ρ in the Sticky-Price and Sticky-Wage Model



only a monetary shock is considered. The findings show that, while the baseline sticky-price model cannot replicate the inflation persistence observed in the data unless an implausible degree of either price stickiness or exogenous nominal interest rate persistence is assumed, a model with both sticky prices and sticky wages can replicate more closely the autocorrelation function of inflation, even with acceptable levels of both price and wage stickiness. This result is important because some notable studies, such as Fuhrer and Moore (1995) and Chari, Kehoe, and McGrattan (1998), have criticized the incapability of this kind of model with nominal rigidities to match inflation persistence.

REFERENCES

Blanchard, Olivier Jean, and Nobuhiro Kiyotaki. 1987. Monopolistic competition and the effects of aggregate demand. *American Economic Review* 77 (September): 647–66.

Blinder, Alan, Edie D. Canetti, David E. Lebow, and Jeremy B. Rudd. 1998. *Asking about prices: A new approach to understanding price stickiness*. New York: Russell Sage Foundation.

Chari, Varadarajan V., Patrick J. Kehoe, and Ellen R. McGrattan. 1998. Sticky price models of the business cycle: Can the contract multiplier solve the persistence problem? Federal Reserve Bank of Minneapolis Staff Report 217, May.

Christiano, Lawrence J., and Martin Eichenbaum. 1992. Liquidity effects and the monetary transmission mechanism. *American Economic Review* 82 (May, Papers and Proceedings of the Hundred and Fourth Annual Meeting of the American Economic Association): 346–53.

Clarida, Richard, Jordi Galí, and Mark Gertler. 2000. Monetary policy rules and macroeconomic stability: Evidence and some theory. *Quarterly Journal of Economics* 115 (February): 147–80.

Dutta, Shantanu, Mark Berger, and Daniel Levy. 1997. Price flexibility in channels of distribution: Evidence from scanner data. Emory University, photocopy.

Erceg, Christopher J., Dale W. Henderson, and Andrew T. Levin. 2000. Optimal monetary policy with staggered wage and price contracts. *Journal of Monetary Economics* 46, no. 2:281–313.

Fuhrer, Jeffrey C., and George R. Moore. 1995. Forward-looking behavior and the stability of a conventional monetary policy rule. *Journal of Money, Credit, and Banking* 27, no. 4, part 1:1060–70.

Galí, Jordi, and Mark Gertler. 1999. Inflation dynamics: A structural econometric analysis. *Journal of Monetary Economics* 44, no. 2:195–222.

Lucas, Robert E. 1976. Econometric policy evaluation: A critique. In *The Phillips curve and the labor markets*, edited by Karl Brunner and Allan H. Meltzer. Supplementary series to the *Journal of Monetary Economics*. Amsterdam: North-Holland.

Rabanal, Pau, and Juan F. Rubio-Ramírez. 2003. Comparing New Keynesian models of the business cycle: A Bayesian approach. Federal Reserve of Atlanta Working Paper 2001-22a, revised February 2003.

Woodford, Michael. 2002. Interest and prices. Princeton University, photocopy.