Corporate Governance and Responsibility

Based on a speech by President Santomero at the Corporate Governance & Responsibility Seminar,
Wilkes University, Wilkes-Barre, PA, February 24, 2003

BY ANTHONY M. SANTOMERO

Corporate governance problems have recently gained the spotlight. What is the root of these problems? The mega-merger frenzy of the 1980s? Overly optimistic forecasts of earnings? Innovations in the financial services industry? Although these explanations are plausible and these factors undoubtedly played some role, President Santomero views governance problems as having deeper roots. The central dilemma is one of conflicting interests in organizations—what economists call "the principal-agent relationship."

Recent headlines have brought the issue of corporate governance to everyone's attention. We have all seen the many stories about Enron, WorldCom, Adelphia, and other companies that were once mainstays of our economy and our business community. Television has brought us images of corporate executives, not being recognized for their civic contributions but being led away in handcuffs on allegations of malfeasance.

A common refrain in all these stories is that company executives were not acting in the best interest of their organizations, their shareholders, and their employees. A combination of inadequate monitoring, a breakdown in internal controls, and the systematic failure of both outside directors and outside auditors appear to have led to a less than desirable outcome.

WHY HAS THIS HAPPENED?

Several reasons have been offered to explain why these corporate governance problems have recently gained the spotlight. Some have argued the origin of these problems was the mega-merger and takeover wave of the 1980s, when innovative compensation programs for top executives were established — including a significant increase in the use of stock options. While these programs were supposed to improve management's incentives to increase shareholder value, some see them as the seeds of our current problems.

These compensation programs expanded and covered more companies during the 1990s. For example, Harvard law professor John Coffee, Jr., points out in a recent paper that in 1990, equity-based compensation for CEOs was 5 percent of total compensation. By 1999, it was 60 percent. Stock options rose from 5 percent of outstanding shares in U.S. companies in 1991 to 15 percent a decade later. Meanwhile, the value of stock options in the largest 2000 companies in the U.S. more than tripled between 1997 and 2000.

Detractors argue that these changes in executive compensation tended to place more emphasis on short-term gains in a company's stock price, rather than on long-term performance. Then, the 1990s brought about rapid changes in technology, greater deregulation, and increased globalization of markets. This placed more pressures on companies' cash flows and made it more difficult to raise share valuations. The innovative compensation programs encouraged executives to take greater risk or to engage in more creative accounting to improve reported earnings. In effect, corporations shifted their business standards and were not held in check by either their corporate directors or others charged with guarding shareholder interests.

Anthony M. Santomero, President, Federal Reserve Bank of Philadelphia

Another explanation of recent events focuses on the fact that long-term earnings forecasts for many companies were overly optimistic during the decade of the 1990s and generated unrealistic expectations. In a speech last year, Fed Chairman Alan Greenspan noted that three- to five-year earnings forecasts for S&P 500 companies averaged almost 12 percent per year between 1985 and 2001. However, actual earnings growth over that period was 7 percent.

Some blame analysts and Wall Street for these overestimates of earnings. They argue financial firms promoted and retained those analysts with the most optimistic forecasts of companies' earnings. Interestingly, the bias to do so was especially pronounced among analysts employed by the underwriting firm.

Still another explanation of these scandals lays blame at the foot of innovations in the field of finance. By this view, these innovations outpaced the ability of traditional accounting and auditing standards to monitor many corporations' activities. They allowed some executives to engineer creative accounting techniques to obfuscate earnings and conceal negative results. Consequently, investors and outside parties had more and more difficulty understanding the financial statements and the risk positions of these large, complex organizations.

Of course, all these suggested explanations received little attention when stock prices were rising rapidly during the bull market. The sharp declines in stock prices have led to greater awareness and concern.

In essence, the foundation of trust was breached between corporations and shareholders with regard to the meaningful disclosure of corporations' financial information. The outcome was a break between executives' pay and the corporations' performance. The whole process of reporting earnings and financial statements became tainted.

AN OLD PROBLEM WITH DEEP ROOTS

Although the problems outlined in these explanations certainly played some role in recent events — or even a major role — focusing on them gives the impression that corporate governance problems are a relatively new issue. I disagree. Rather than a recent development, such issues have deep roots. They are inherent in what economists call "the principal-agent relationship" in organizations.

> The oversight of the firm falls directly on the company's board of directors.

The central dilemma here is one of conflicting interests. Much research has been devoted to how to provide incentives to the agent — or executive management of a firm, for the purposes of our discussion — to act in the best interests of the principals — the owners of the firm. In essence, the challenge of our form of capitalism is, and has always been, to construct a system of corporate governance so that company management acts in the best interests of shareholders.

This is what we have attempted to do in our structure of corporate governance. The oversight of the firm falls directly on the company's board of directors. In the end, the board bears ultimate responsibility for the company's performance. The board is supposed to implement methods to monitor and control management so that abuses are prevented or at least minimized. To do this, some members of the board of directors are outsiders — people who are not part of the management team of the company. These directors are supposed to act as an independent check on corporate management to ensure they act in the shareholders' best interest. American capitalism relies on the fiduciary concept to protect those who entrust their money to large — and often distant — corporations.

THE IMPORTANCE OF CORPORATE GOVERNANCE

Yet today, there is a sense that the model just described is not working well enough. A crisis of confidence in corporate America has resulted. Recent scandals have generated a lingering sense of uncertainty and vulnerability among investors. This has put pressure on companies' management, corporate directors, and regulators to address problems of accountability and control.

To restore public confidence in the integrity of corporate America, companies must demonstrate a strong commitment to the development and enforcement of rigorous standards of corporate governance. These standards must encompass the relationship between a company's board of directors, its management, and its shareholders. They must require corporate leaders to be faithful to shareholder interests and act with both competence and integrity.

At a very basic level, trust is at the heart of the free enterprise system. But the current state of public trust in American corporations is not good. According to a recent poll, 77 percent of the public believes CEO greed and corruption caused declines in the stock market. In addition, polls suggest much of the public rejects the view that the scandals were isolated incidents within a system in which most corporate leaders are good and honest people.

Yet, the continued success of our economic system requires the confidence and trust of investors, employees, consumers, and the public at large. In short, there is much work to be done.

WHAT IS BEING DONE?

We know that as corporations grow larger and more complex, it becomes more difficult for boards of directors to monitor activities across the company. Directors cannot be expected to understand every nuance of or oversee every transaction. They should look to management for that.

Nonetheless, the role of a corporate board of directors is quite substantial, and directors are required to be highly knowledgeable. They must know — understand — the nature of the firm's business, its financial performance, and the nature of the risks facing the firm's strategic plan. Collectively the board should have knowledge and expertise in areas such as business, finance, accounting, marketing, public policy, manufacturing and operations, government, technology, and other areas necessary to help the board fulfill its role. They must set the tone for risk-taking in the institution and establish sufficient controls so its directives are followed. They also have the responsibility to hire competent individuals who possess integrity and the ability to exercise good judgment. Members of the audit committee, in particular, must be independent and have knowledge and experience in auditing financial matters. This is no small task.

Recent events have been a loud wake-up call, focusing attention on the need to heighten our commitment to proper corporate governance and improve both accountability and control. In response, a number of measures have been taken or proposed by various groups to bolster confidence in our corporate system.

Recognizing that boards have come under increased scrutiny, the New York Stock Exchange has appointed a Corporate Accountability and Listing Standards Committee. The committee has come up with a number of recommendations to improve corporate governance. One proposal is to increase the role and authority of independent directors by having them make up a majority of a company's board. The committee also recommends that companies adopt corporate governance guidelines and a code of business ethics and conduct. In addition, the committee has suggested shareholders be given more opportunity to monitor the governance of their companies. They must vote on all equity-based compensation plans and have access to the company's corporate governance guidelines.

The U.S. Financial Accounting Standards Board (FASB) is considering how to improve accounting standards for special-purpose entities.

The Conference Board Commission on Public Trust and Private Enterprise has also made recommendations on best practices. It recommended that executive compensation be performance-tied and stressed the importance of independent directors being able to retain outside consultants. The commission also suggested that the Federal Accounting Board (FAB) and the International Accounting Standards Board (IASB) come up with standardized definitions of revenues in order to achieve true parity in determining executive compensation based on company performance. Finally, it recommended that America's senior executives should be subject to much longer-term holding periods for company stock and higher ownership requirements.

A consensus is now also growing concerning some needed changes to certain underlying accounting standards and their application. The U.S. Financial Accounting Standards Board (FASB) is considering how to improve accounting standards for special-purpose entities. This is in response to the growth of securitization and the added complexity securitization has introduced into financial reporting.

A pilot program is under way to standardize financial reporting data and make them available to investors via a web site hosted by Nasdaq. In the future, technology will be instrumental in improving transparency in financial reporting by making corporate financial information easily available.

Congress has also responded. The newest legislation about corporate governance, the Sarbanes-Oxley Act, addresses the wave of recent events that shook public confidence. The act seeks to protect investors by improving the accuracy and reliability of corporate disclosures. Among its major provisions is the establishment of a new private-sector regulatory regime in which the SEC handpicks an oversight board to monitor standards and conduct in the accounting industry. In fact, my colleague Bill McDonough, president of the New York Fed, has been tapped to head this new board.

Sarbanes-Oxley also emphasizes the need for a wall of independence between auditors and firms. In addition, and perhaps most controversially, the act seeks to strengthen corporate responsibility by creating a structure for holding individuals and companies criminally and/or civilly accountable for their actions. CEOs and CFOs are now required to certify quarterly and annual reports to ensure proper disclosures.

While some may disagree with any of these proposals, it is important to realize that those involved and responsible have begun to take action to address the perceived problems of corporate governance.

CORPORATE GOVERNANCE IN BANK REGULATION

The nonfinancial sector is not alone in its search for better corporate governance. The requirement of trust and confidence in corporate America is analogous to the trust and confidence issues that the Federal Reserve faces in its role as the regulator of the U.S. banking system. Let me touch on some of the parallels and briefly describe how we have addressed them.

Of course, banks have shareholders, too. Their business involves making loans to customers who are expected to repay. Bank management has a good deal of information about the quality of the loan portfolio. The question facing bank managers and their directors is how much information to provide to shareholders. As stewards of the public trust, bank regulators and supervisors ask the same questions.

But beyond this, a bank's relationship with its depositors is another example of the principal-agent problem. Depositors depend upon bank regulators and supervisors, as well as deposit insurance, to keep their money safe in spite of the opaque nature of bank assets. As a result, we have substantial interest in the ways in which corporate governance is performed in the regulated banking sector, and we have incorporated these concerns into our regulatory and supervisory model.

The primary focus of the Federal Reserve's approach to supervision and regulation is ensuring an institution's safety and soundness. The Federal Reserve's examiners also ensure compliance with banking laws and regulations, including consumer-protection laws and regulations.

Historically, a major focus of banks and their regulators has been on whether they accurately report their financial condition and appropriately assess the quality of their assets. Beyond this, supervisors have long been concerned about the quality of internal controls. During the past 15 years, the Fed's supervisory program has been broadened to focus on banks' overall risk-management systems, comparing them against both regulatory standards and industry best practices.

In 1991, Congress broadened the scope of banks' assessments of risks and controls.

The Fed's risk-assessment process analyzes the nature and extent of risk to which a financial institution is exposed and assesses how well the institution is identifying, controlling, and managing risks. It requires integrated, enterprisewide risk management that considers all areas of risk, including credit risk, market risk, liquidity risk, operational risk, legal risk, and reputational risk. The idea is to identify not only the type of risk and its level but also its direction and whether the bank has means to effectively control each risk.

The Fed also wants to ensure that the bank has a strong internal audit function and that it also receives a thorough, complete, and independent external audit. To accomplish all this, Fed examiners conduct on-site examinations and provide institutions with continual off-site monitoring and analysis as economic conditions and the bank's financial condition change.

In 1991, Congress broadened the scope of banks' assessments of risks and controls. Since then, bank managers are required, at least annually, to step back from other duties and evaluate risks and internal controls. In addition, external auditors must attest to management's results of this self-assessment of risks and internal controls. The results are reported to the audit committee of the bank's board of directors. Incidentally, the audit committees of banks' boards have been required to be independent of management for a long time — something that is now being stressed for all corporations. In fact, this approach to risk-assessment and internal controls is also the one followed by all of the Federal Reserve Banks for several years.

Ensuring a broad-based assessment of risks and internal controls has served the banking industry well in recent years. For instance, despite the economic downturn in 2001, most banks continue to be in good health.

BUT THERE ARE LIMITS

The Fed's experience, therefore, suggests some success with the evolving model of better corporate governance. Nonetheless, it is important to remember that the process is still evolving. Much work remains to be done. It is important to remember that proposals to improve corporate governance must take into account not only the expected benefits of new standards and regulations but also their expected costs to both the corporations and the economy as a whole. Good intentions do not always prevent unintended consequences. Some unintended side effects might include high compliance costs, ambiguous liabilities, or reduced innovation. This is particularly true when various proposals about corporate governance have been arising at both the state and federal levels.

Disclosure should never be so onerous as to make the cost of compliance prohibitive or impractical. Regulations and standards of any sort — whether by regulators, government, trade groups, or the companies themselves — should not excessively impede the ongoing process of innovation. Rather, we must ensure an environment conducive to markets that are effective and efficient, safe and sound.

IS THERE A BETTER WAY?

One criticism of the past approach to corporate governance is that it tended to focus on the development of fairly specific rules of behavior rather than insisting on adherence to certain principles of behavior. Most of the proposals we now debate to improve corporate governance are new rules.

However, the problem with rules, particularly accounting-based rules, is that innovations in the financial system can open loopholes in rules. When loopholes open, inappropriate or unethical actions that are not specifically prohibited by the rules can take place. Basically, this is what happened in many of the corporations that made news headlines in the past year or two.

A credible case can be made that we should focus on principles instead of rules. That is, we should establish key principles against which corporate decisions should be held accountable, regardless of whether a certain type of behavior is prohibited. This way, when innovations make old rules obsolete, corporate leaders and their financial executives would have to consider not only whether some action would violate a rule but whether it would violate a principle.

There are strong arguments for developing principles-based standards — in addition to our reliance on traditional rules-based standards. However, the challenge is to establish a set of principles that are sufficiently clear and concise. This may not be an easy task, and the result may be substantial litigation, rather than simplification and clarity. It is important to consider both the costs and the benefits of new standards, as well as the unintended consequences that may result. In the end, rules cannot replace ethics and an exemplary “tone at the top.”

Nonetheless, whichever way regulation evolves, disclosure and transparency are imperative to adequate corporate governance. Such disclosure need not be identical across all industries and companies. The information available to the public should be what is necessary for them to evaluate a particular firm’s risk profile. That is why principles-based accounting has some appeal. Companies should take action to ensure their financial statements divulge what is truly essential for investors to understand the business and make informed decisions.

Companies should take action to ensure their financial statements divulge what is truly essential for investors to understand the business and make informed decisions.

CONCLUSION

Good corporate governance is critical to the health of the corporate system, our financial system, and our economy. Our economy will be stronger if corporate decisions are made with competence and integrity, and if shareholders and the public can appropriately assess the profitability and riskiness of corporations’ business activities.

The crisis of confidence in corporate America has been created by recent scandals that have generated a sense of uncertainty and vulnerability among investors. These events have put pressure on regulators, corporate directors, and management to address problems of corporate accountability and control. Changes are in the works and appear to be in the right direction. To a large extent, this direction is where banks and bank regulators have gone before.

Many ideas to improve corporate governance are being offered by a variety of parties. Adopting a system of principles-based accounting standards, rather than primarily amending the current rules-based standards, may be useful in ensuring that our accounting rules do not become quickly out of date in the face of rapid financial innovation. But given the wide range of ideas being offered, the challenge will be to move forward and implement those proposals most likely to be effective in yielding benefits at a reasonable cost of compliance and to do so without generating unintended consequences. The challenge is in the implementation, but the challenge is a noble one. We must proceed.

In the end, however, we must bear in mind that the core principles of ethical behavior and sound business practices are the keys to any real success in this arena. This tone is set at the top. Without these values we will never really succeed in conquering the problems and conflicts that arise in corporate governance.

U.S. Coins: Forecasting Change

BY DEAN CROUSHORE

Although the government annually produces about 70 new coins for every man, woman, and child, the economy's need for coins can vary from year to year. So how do the U.S. Mint, which makes the coins, and the Federal Reserve, which distributes them, decide how many coins the economy needs? Dean Croushore highlights some facts about coins and describes how demand for change is forecast.

Every year, the U.S. government produces about 70 new coins for every man, woman, and child in the country, or about 20 billion coins. In recent years, the program to produce new state quarters, plus the introduction of the golden dollar, increased total demand for new coins. When the demand for the new quarters and dollars became surprisingly strong in 1999 and 2000, shortages of some coins developed in different parts of the country.

To help prevent such shortages, a team of economists and analysts at the Federal Reserve are developing new models to forecast demand for coins. This article describes some of the work the Philadelphia Fed has undertaken since the project began in 2000.

Dean Croushore is vice president and economist in the Research Department of the Philadelphia Fed.

Let's see how the Federal Reserve and the U.S. Mint decide on the number of coins to be produced and distributed and how difficult it is to forecast demand for coins. We'll begin by looking at some basic facts about the institutions involved and how demand for coins is calculated.[1]

THE FACTS ABOUT COINS

The U.S. Mint is in charge of producing coins. The Mint must obtain the raw metals to be used in production, procure the equipment, and hire workers to produce the coins needed for the economy. Once coins are produced, the Mint sells them at their face value to the Federal Reserve. The Mint makes a considerable profit (called seignorage) on the sales of coins. For example, one of the new Sacagawea golden dollars costs about 12 cents to produce and yields one dollar in revenue. The resulting profit of 88 cents goes to the U.S. Treasury. The Mint's profits increase the government's revenue by a billion dollars or more each year.

The Federal Reserve distributes coins from 37 coin offices, mainly Reserve Banks and their Branches, and more than 100 coin terminals operated by armored carriers such as Brinks and Loomis-Fargo. The armored carriers hold inventories of coins for the Fed and for banks and other financial institutions (hereafter just called banks). The carriers move coins between their terminals, banks, and Federal Reserve coin offices.

The U.S. Mint generally produces coins for circulation according to orders from the Federal Reserve. The Federal Reserve, in turn, generally orders an amount of coins based on the expected demand from banks. And, of course, those banks want coins to satisfy the demands of their customers. So, ultimately, the amount of coins produced depends on the demand for coins by people and businesses. Let's take a look at how that demand is measured.

The main concept in analyzing the demand for coins is called *net pay*. Net pay is an unusual economic concept because it represents the change in banks' demand for coins, which can be either positive or negative: It's positive

[1] The author thanks David Griffiths of the U.S. Mint for comments on an earlier draft of this article, and Brian Coulter of the Federal Reserve Cash Product Office, Geoffrey Gerdes of the Board of Governors, and Christopher Sims of Princeton University for consultation on the coin forecasting models. The author further thanks the team of economists and analysts at the Federal Reserve Bank of Philadelphia who worked on this project, including John Chew, Brian DiCiurcio, James Gillard, James Sherma, Keith Sill, and Tom Stark.

when banks ask the Federal Reserve to deliver coins because demand has increased; it's negative when those institutions return coins to the Federal Reserve because demand for coins has declined.[2] Because of the durable nature of coins and because they can be returned to the Federal Reserve, coins (as well as paper money) are different from all other goods and pose unique challenges.

Coins flow between people or businesses and banks (Figure 1). When people and businesses want more coins, their banks order more from their local Federal Reserve offices, which then ship coins to the banks, either from inventories of coins located at armored carriers (the line labeled *A* in the chart) or directly from the offices' own inventories (line *B*). Occasionally, the U.S. Mint ships coins directly to a bank (line *C*). Each of these shipment methods to banks results in positive net pay, since banks' demand for coins is positive.

However, if people begin turning in more coins than banks want to keep on hand, the banks may return the extra to the Federal Reserve, either directly (line *D*) or through armored-carrier terminals (line *E*). So lines *D* and *E* represent negative net pay because demand for coins is negative.

In any given month, some banks may have positive net pay and others may have negative net pay. Net pay is the sum of the amounts shown by lines *A*, *B*, and *C* minus the sum of the amounts shown by lines *D* and *E*:

$$Net\ pay = (A+B+C)-(D+E).$$

Net pay is calculated separately for six different denominations of coins: penny, nickel, dime, quarter, half-dollar, and dollar. Total net pay is the sum across all banks.

Let's take a look at the data on net pay for each denomination to see how each has changed over time. In doing so, we will look at the net pay for each denomination in units of millions of coins each month. The data run from 1957 to 2002, except for half-dollars and dollars. For each denomination, the black line shows the actual monthly amount of national net pay (summed across all 37 Federal Reserve offices). The green line is the average volume over the past 12 months (called a 12-month moving average), which is shown to help illustrate the long-term trend in the data (Figures 2a to 2f).

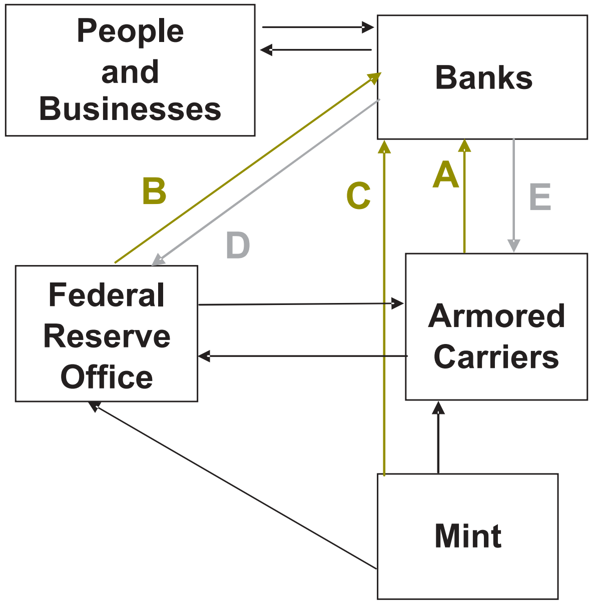
The charts show some interesting patterns. First, you can see that month-to-month seasonal fluctuations in net pay are huge. For some denominations, national net pay is even negative in some months. The net pay of different denominations swings dramatically from one month to the next, mostly because of changes in people's spending patterns. We need a lot more change in the summer months for parking meters at the shore and for soda machines. We use more change around holidays at the end of the year, as well. But we need much less in the middle of the winter.

About half of all coins produced are pennies. Net pay of pennies has averaged over 800 million coins per month in the last five years, while the sum of all other denominations has been about 700 million coins per month. The net pay for pennies has been fairly constant since 1980, perhaps trending down slightly (Figure 2a). For other main denominations (nickels, Figure 2b; dimes, Figure 2c; and quarters, Figure 2d), the trend over time has been slightly upward, which suggests that these other denominations may be gradually replacing pennies in terms of quantity used for making change.

There are some interesting variations in net pay for those coins, especially in the 1960s when the value of silver, which was a major component of dimes and quarters, increased sharply. Demand for those coins declined

FIGURE 1

Net Pay



[2] Note that net pay differs from the change in the demand for coins in the rare instance in which there is a shortage of coins. Because such cases are rare and we do not have reliable data on the amount of shortages, our models of coin demand ignore such instances and assume that net pay equals the change in demand.

sharply when the coins were redesigned with no silver in them. The other two denominations, half-dollars and dollars, had net pay near zero for much of the 1980s and 1990s. The introduction of the Sacagawea golden dollar in 2000 caused a sharp increase in the net pay of dollars that year.

Given these trends in demand for different denominations, what can we say about overall demand for coins? To investigate this issue, we'll examine total demand for coins in terms of numbers of coins, adding up the net pay for all six denominations. Also, to avoid confusion arising from seasonal fluctuations, we just look at the total net pay in each calendar year (Figure 3). We look at the total number of coins rather than their dollar value, in part because the Mint's ability to produce enough coins to meet demand depends on the number of coins rather than their dollar value.

In the graph, you can see that overall net pay generally increased over time, from under 2 billion coins in 1957 to a peak of 23 billion in 1999 and 2000. But the increase was not steady. From one year to the next, sometimes net pay rose and sometimes it fell.

We might expect a correlation between net pay and the strength of the economy because it seems likely that people will use more coins if they're buying more goods and services. However, there does not appear to be a strong correlation between net pay and economic activity. For example, while net pay fell when the economy weakened, as in 1990 and 1991, it fell even when the economy was strong, as it was in 1996 and 1997.

Special events raised net pay to very high levels in 1999 and 2000. First, beginning in 1999, the Mint (directed by laws passed by Congress) rolled out the first quarters in the state commemorative program. The demand for these new quarters turned out to be significantly stronger than anticipated, thus

FIGURES 2a - 2c

Figure 2a
Pennies Net Pay

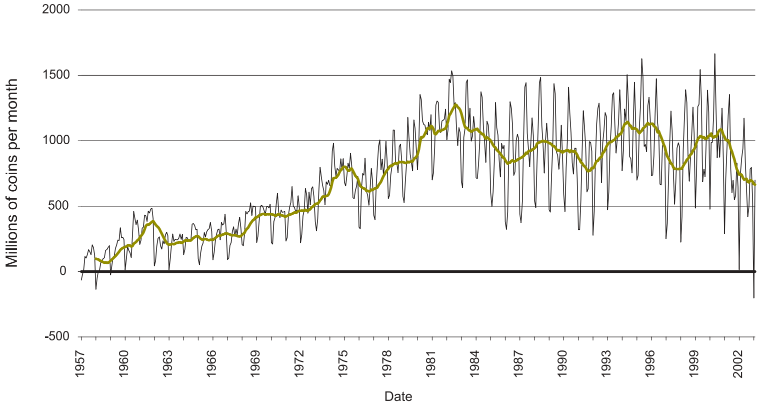


Figure 2b
Nickels Net Pay

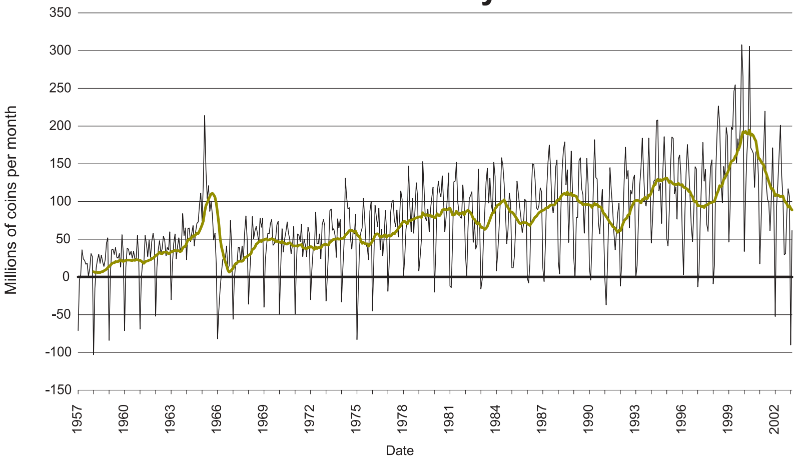
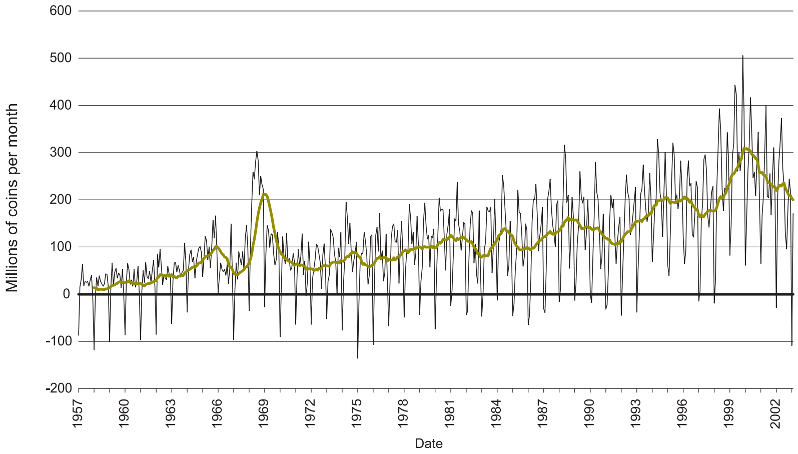


Figure 2c
Dimes Net Pay



Note: The black line shows the monthly value of net pay; the green line indicates the 12-month moving average.

causing net pay to rise sharply (Figures 2d and 3). Then, in 2000, the Sacagawea dollar was introduced to much fanfare. Initial demand for the new coin was also strong, and the Mint produced over 1 billion of them. At the same time, the demand for the new state quarters increased 50 percent from the year before, so again net pay was much higher than expected. At the same time, the demand for nickels and dimes also rose substantially (Figures 2b and c).

FORECASTING COIN DEMAND

Sharp, unexpected increases in net pay during 1999 and 2000 led the Federal Reserve Bank of Philadelphia to investigate ways to improve forecasts of demand for coins. Developing a forecasting model involves several steps: choosing among different types of models, testing the different models to see how well they perform, seeing how they deal with changes, such as the introduction of the new quarter and dollar coins, then running forecasts in real time and investigating the quality of the forecasts. Because demand for each coin denomination seems to behave differently from that of the other coin denominations, the models we examine will contain a separate forecasting equation for each denomination, rather than modeling overall coin demand in a single equation.

Four Models of Coin Demand. We considered four different types of models: (1) a structural model; (2) a time-series model; (3) a vector autoregression (VAR) model; and (4) a Bayesian vector autoregression model. Brief descriptions of each model follow.[3]

Structural Model. Adapting the work of earlier researchers who had modeled coin demand, we first exam-

FIGURES 2d - 2f

Figure 2d
Quarters Net Pay

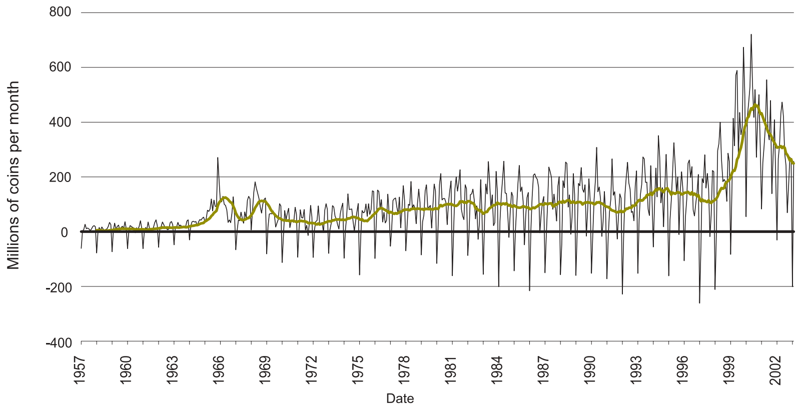


Figure 2e
Halves Net Pay

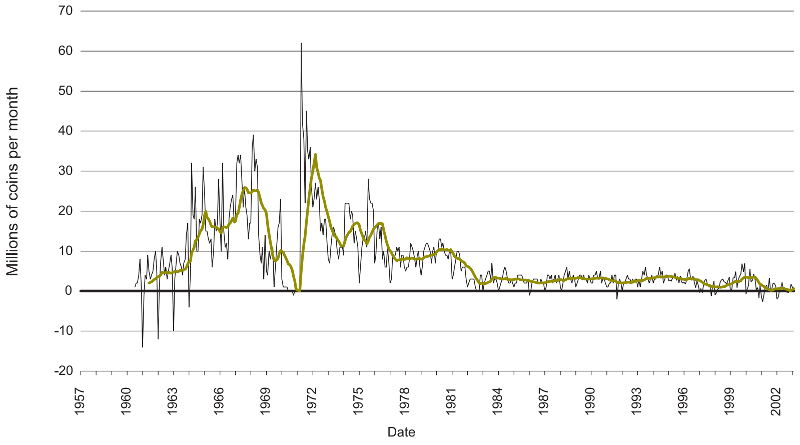
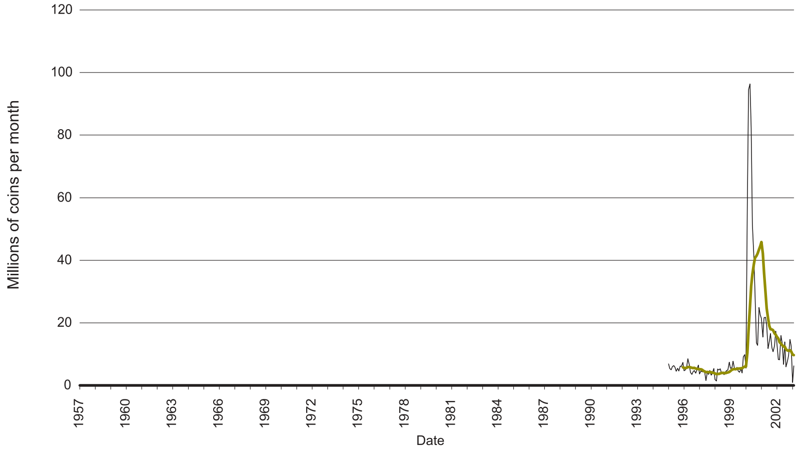


Figure 2f
Dollars Net Pay

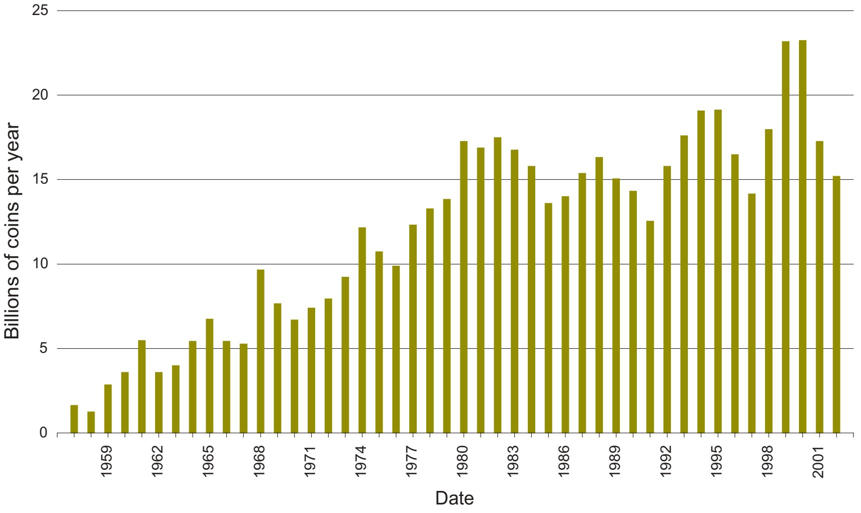


Note: The black line shows the monthly value of net pay; the green line indicates the 12-month moving average.

[3] Additional details about each model can be found in the research paper that I wrote with Tom Stark. The paper is listed in the References section at the end of this article.

FIGURE 3

Annual Net Pay of Coins



ined a structural model of net pay for each denomination. In a structural model, economic theory dictates which variables should affect demand for each denomination. We then develop a forecasting equation based on that economic theory.

Because many economic data are quarterly, we took quarterly averages of the monthly coin data. Economic theory suggests that demand for coins depends on economic activity, interest rates, and the inflation rate. We experimented with a number of measures of economic activity, including consumer spending on services (used in older models), retail sales, industrial production, personal consumption expenditures, and payroll employment. Payroll employment gave the best results in our tests, so we used that variable to represent economic activity. For an interest rate, we used the federal funds rate, which is the interest rate that banks charge each other on overnight loans. Since the fed funds rate is also the main variable the Federal Reserve targets with its monetary policy, it is a good indicator of the overall level of short-term interest rates. For inflation, we chose the inflation rate as measured in the Consumer Price Index. Again, based on an older model, the inflow of coins from banks to the Federal Reserve is modeled separately from the payout of coins from the Federal Reserve to banks. The forecast for the inflow is then subtracted from the forecast for the payout to generate a forecast for net pay. The forecasting model for each coin denomination also includes a seasonal variable for each quarter of the year to account for the seasonal pattern in coin demand.

Time-Series Model. The second forecasting method we tried was a time-series model, which uses data only from the past and data only on the variable being forecast. For example, the model for net pay of pennies assumes that net pay of pennies in the future depends only on past movements of net pay for pennies; it does not depend on the net pay of any other coin denomination or on any macroeconomic variable.

Though such a model is very simple, we still had to make choices about the forecasting equation of the model: how far back to go in determining the forecast, whether to model the level of net pay or the change in the level of net pay from one month to the next, and how to deal with the seasonal fluctuations — seasonally adjust the data before they go into the model or account for seasonal fluctuations within the model. Experimentation suggested that the best model was arrived at by using 14 months of lagged data for each forecast, modeling the change in net pay from one month to the next, and adjusting the data for monthly fluctuations before running the forecasting equation.[4]

Vector Autoregression (VAR) Model. In the past 20 years, many economists have stopped using structural models for forecasting because these models require more precise economic theory than we usually know; economists have also moved away from time-series models because such models use no economic theory at all. A vector autoregression (VAR) model is a mixture of a structural and a time-series model. The VAR uses economic theory to tell the researcher which variables should be included in the model, but it also incorporates time-series techniques by including past data on each variable in the model. A VAR is useful because it allows us to conduct “what-if” experiments, such as: “What will happen to demand for coins if the economy goes into a recession?”

For our VAR model, the net pay of each coin denomination depends on past data on economic activity, the interest rate, and the inflation rate, just as in the structural model. But the equation for each coin denomination depends on that denomination’s own history, just as in the time-series model. In the VAR, the economic data

[4] The best model is determined on the basis of the root-mean-squared forecast error, which is described in detail later.

influence forecasts for coin demand, but coin demand is not allowed to influence forecasts of the economic data. Experimentation showed that the best model came from using data on the logarithm of payroll employment as the variable related to economic activity; that variable proved better in our tests than the growth rate of employment and was also superior to other variables, including industrial production, retail sales, and personal consumption expenditures. We also found it was best to use employment data that were not seasonally adjusted but to account for seasonal fluctuations within the forecasting model. Using 13 months of past data also provided the best results for this model.

Bayesian VAR. Bayesian techniques basically involve a researcher's beliefs (for example, concerning seasonality) about the outcome of an empirical investigation: The researcher examines the data in light of those beliefs, then sees if his beliefs change after he has examined the data. Essentially, Bayesian techniques help us rule out certain outcomes that differ so much from economic theory that we do not believe them. The techniques help to keep the estimated forecasting model within certain bounds. Economists use Bayesian methods because research has found that such methods often generate superior forecasts and may handle monthly fluctuations in the data better. Because of the large seasonal fluctuations in the coin data, Bayesian techniques may be very fruitful.

To implement Bayesian methods for forecasting net pay for coins, we applied them to the VAR model described above. The main differences between the VAR and the Bayesian VAR are in the amount of past data used (24 months in the Bayesian version versus 13 months in the non-Bayesian version) and in the coefficients of the forecasting equation, which are fixed in the VAR but allowed to change over time in the Bayesian VAR. In the Bayesian VAR, some key coefficients are chosen to make the model perform well, the most important being those that concern the seasonal patterns in the data. The Bayesian VAR is also not as restrictive as the VAR because it allows the data on coins to affect the macroeconomic variables (perhaps because people's spending habits are reflected in coin demand, which then helps predict the macroeconomic variables) and it allows data on one coin denomination to affect the forecasts for other coin denominations.

Because of the large seasonal fluctuations in the coin data, Bayesian techniques may be very fruitful.

Comparing the Models. After we built the models, we tested their performance over several periods. To avoid being unduly influenced by the introduction of the new state quarters and by the new dollar coin, we chose 1990 to 1998 as our main testing period. To see how the models would have performed over that time, we generated forecasts at the start of each three months, as if we were at that date and did not know what was to come. That is, we first used the coin data from January 1957 through December 1989, which would have been known to a forecaster making a forecast in January 1990, and generated a forecast for the next 12 months for each coin denomination. Then we stepped forward three months, as if we were in April 1990. Then, using the coin data through March 1990, we generated a forecast for the next 12 months. We continued this process until January 1998, at which point we generated forecasts through the end of 1998 (just before the start of the new state quarters program).

With these forecasts in hand, we calculated summary statistics on how well each forecast did. The most appropriate statistic is the root-mean-squared forecast error (RMSFE), which quantifies deviations — either positive or negative — of the actual from the forecast, where larger errors are penalized more. The RMSFE is calculated by taking a forecast for each year, calculating the forecast error (actual minus forecast) for each month, squaring each error, adding up the squared errors, dividing by the number of forecasted values, then taking the square root. Researchers have found that the RMSFE has a number of desirable properties and gives them a general guide to using forecasts: The best forecasts are those with the lowest RMSFEs. By squaring the forecast errors in calculating the RMSFE, forecasts that are far from actual are penalized more heavily than if we just calculated the average error (Table 1).

As you can see from the table, the time-series model, which was originally proposed as a benchmark model, proved very difficult to beat. In fact, only the Bayesian VAR did better, and its improvement was only slight.[5]

Using these models, we began to generate forecasts periodically, as requested, first by the Fed's Cash-Fiscal Product Office (located at the Federal

[5] Of course, all the forecasting models did much worse in forecasting coin demand in 1999 and 2000 because nothing in the models accounted for the introduction of new coins. But a researcher could have used these models to forecast net pay for purposes of determining how many coins were needed for circulation, then added a projection for demand for new coins that would not circulate because people would keep them as collector's items.

Reserve Bank of Philadelphia), and then the Fed's Cash Product Office (located at the Los Angeles Branch of the Federal Reserve Bank of San Francisco) when that office took over the responsibility for coin issues in spring 2001. Because the structural model performed so poorly in our tests, we stopped generating forecasts with it in early 2001. Instead, we added the Bayesian VAR to our process in September 2001.

Now, after the Federal Reserve coin offices calculate data on net pay at the end of each month, we generate new forecasts for net pay at the national level using the time-series, the VAR, and the Bayesian VAR models. Because no one knows how long or how large the increased demand for state quarters or the demand for the new dollar is likely to be, the best forecast is likely to be one that simply tracks the overall trend but does not generate forecasts that make strong assumptions about that future demand. All the models we use have that feature. For example, in 2001, demand for coins slowed substantially. Although the models did not predict the slowdown, the forecasts adjusted fairly quickly after the slowdown began.

How Have the Forecasts Performed So Far? The key question for any forecasting method is: How well does it work? Unfortunately, we have been forecasting demand for coins only for about two years, so we cannot answer that question very well. Table 2 shows the forecasts made every three months from February 2001 to November 2002, along with the actual values in 2001 and 2002. As you can see in the table, the initial forecasts for 2001 and 2002 were fairly high. Given what had happened in 1999 and 2000, with coin demand rising, the forecasting models predicted continued strong demand in 2001 that did not materialize. Instead, coin demand began declining substantially, and it took the forecasting models several months to adjust fully.

So far, it appears that the time-series model has done the best job of forecasting because it was the quickest to lower forecasts for 2001 and 2002 as net pay fell. But the period is much too short to favor the use of that model over the others. In a few years, we will have much more data on the forecasts and the errors made by each model, and we will be able to undertake a more complete examination.

USING THE FORECASTS

How can the Federal Reserve use these forecasts for coin demand? First, the national coin forecasts can help the Mint in planning its production. The Mint needs to schedule workers and to purchase enough equipment to produce the right amount of coins. Improved forecasts will allow the Mint to reduce production costs by getting a better idea of how many coins it will need to produce. In addition, the Mint will be able to order the appropriate amount of raw materials needed for production.

The forecasts can also help the Federal Reserve in ordering coins. Each Federal Reserve office must order coins every month, and improved forecasts could give them an additional tool for deciding how much to order. Those offices maintain inventories of coins in case of sudden changes in demand, so improved forecasts can help them maintain appropriate levels of inventories. Improved forecasts can also help these offices reduce costs by keeping inventories from becoming too large or too small, since shipping coins between offices is costly.

Because the time-series models performed so well at the national level, we began forecasting net pay for each office based on such models in spring 2002. Because there are 37 offices and six coin denominations, we generated 222 forecasts (37 x 6), each running monthly for the next 30 months. These forecasts are distributed to each office for its use in ordering coins and for planning. Coin offices must also take into account changes in local and national economic conditions that may not be captured in the time-series model that forecasts net pay.

SUMMARY

Forecasting the demand for coins is difficult because of seasonal

TABLE 1

RMSFE for Different Coin Models

Model	RMSFE
Structural Model	2.61
Time-Series Model	1.75
VAR	2.01
Bayesian VAR	1.72

Note: Figures shown are the root-mean-squared forecast error (RMSFE) for each model over the testing period from 1990 to 1998, in billions of coins. A forecast is more accurate if it has a smaller RMSFE. Models for half-dollars and dollars were not run because we have insufficient data.

fluctuations in net pay and the introduction of new coins. By using some standard types of forecasting models, we have attempted to improve on existing forecasts of net pay. Whether these forecasting models will perform well in practice will require several years of observations. The models we use depend on the stability of historical relationships. As such, changes in how people use coins could cause the models to make large forecast errors in the future. If the models do not do well, we may be able to modify them so that they forecast better in real time.

Our hope is that we will be able to forecast coin demand well enough to prevent any shortages of coins in the future, without the expense of piling up large inventories of unused coins.

TABLE 2

Annual Real-Time Net Pay Forecasts

Forecast Date	Actual Data Through	Calendar Year Forecasts	
		2001	**2002**
Feb 2001	Jan 2001	21.4	21.7
May 2001	Apr 2001	20.8	20.5
Aug 2001	July 2001	18.1	18.1
Nov 2001	Oct 2001	16.7	15.2
Feb 2002	Jan 2002		14.0
May 2002	Apr 2002		17.1
Aug 2002	July 2002		17.0
Nov 2002	Oct 2002		
Actual		**17.1**	**15.2**

Note: Amounts in billions of coins per calendar year.
Numbers shown for forecast dates from February 2001 to August 2001 are the average forecasts from the time-series model and VAR; numbers shown from November 2001 on are the average forecasts from the time-series model, VAR, and Bayesian VAR. Each forecast is a projection for the calendar year shown in the column header for the last two columns.

REFERENCES

Croushore, Dean, and Tom Stark. "Forecasting Coin Demand," Federal Reserve Bank of Philadelphia Working Paper 02-15/R, September 2002.

Roseman, Louise L. "The Golden Dollar Coin: Testimony Before the Subcommittee on Treasury and General Government of the Committee on Appropriations, U.S. Senate," May 17, 2002.

United States Mint, *2001 Annual Report*.

Antitrust Issues in Payment Card Networks:
Can They Do That? Should We Let Them?

BY ROBERT HUNT

In the United States, payment card networks coordinate the activities of thousands of financial institutions that issue cards, millions of retail locations that accept them, and several hundred million consumers that use them. This coordination may include the collective setting of certain prices and other controversial network rules. Such practices have recently come under the scrutiny of antitrust authorities in the U.S. and abroad. In this article, Bob Hunt describes the economics of the payment card industry and explains how it differs from the textbook model of competitive markets. He argues that these differences should be reflected in the antitrust analysis of payment card networks.

In the United States, general-purpose payment cards — Visa or MasterCard, ATM cards, or debit cards — are ubiquitous and easy to use. In 2000, there were about 900 million general-purpose payment cards in the U.S., or about four for every adult. These cards were used in 28 billion transactions worth $1.9 trillion. Indeed, payment cards are displacing the paper check at the point of sale — the number of consumer checks written peaked during the 1990s and is now in decline.[1]

In this article, I explore how payment cards work and explain why we need to think a little differently about the market for consumer payment methods than we do for most other markets. This has implications for when, why, and how the rules of antitrust law — which regulate how firms may exercise market power — should be applied to this industry. This is not just an academic question: In the U.S. there are currently two important antitrust cases involving payment cards. Australia recently introduced new regulations for the payment card industry in that country while the U.K. and the European Union have contemplated similar measures.

Payment cards have two distinguishing features that lead us to think differently about this market. First, payment cards exhibit what economists call network externalities — for example, payment cards are more valuable to consumers when more merchants accept them. Second, in the U.S. at least, thousands of banks and other firms provide payment card services to millions of cardholders and millions of merchants who accept cards. In an environment with network externalities and so many participants, economic theory suggests that some form of coordination is beneficial, possibly essential. In the U.S., this is usually done by forming a payment card network.

Payment networks coordinate the behavior of banks, merchants, and consumers by setting certain prices and rules. In many other contexts, such practices might be considered anti-competitive. It is also possible they can have anti-competitive effects in the market for consumer payments. Yet a careful examination of economic theory tells us this is not always the case.

The challenge to policymakers is to decide, based on the available information, whether a network's

Bob Hunt is an economist in the Research Department of the Philadelphia Fed.

[1] These statistics are from the Bank for International Settlements' *Statistics on Payment Systems in Selected Countries* (2002), Table 6. They exclude store cards. The decline in checks' share of consumer transactions — relative to credit and debit card transactions — is documented in the article by Geoffrey Gerdes and Jack Walton.

pricing strategy and rules are likely to advance or retard economic efficiency. Such conclusions are complicated by dynamic considerations — a network that exercises market power may spur the development of competing networks with superior technology.

THE ORGANIZATION AND ECONOMICS OF PAYMENT CARD NETWORKS

The U.S. payment card industry involves thousands of banks participating in a number of networks, millions of consumers who find it valuable to use a payment card, and millions of merchants who find it valuable to accept those cards.[2]

Pricing and Rulemaking in Payment Card Networks. Banks engage in two types of activities within a payment card network.[3] Card *issuers* are banks that offer cards to consumers and determine the level of any fees or finance charges their customers see on their regular statements.

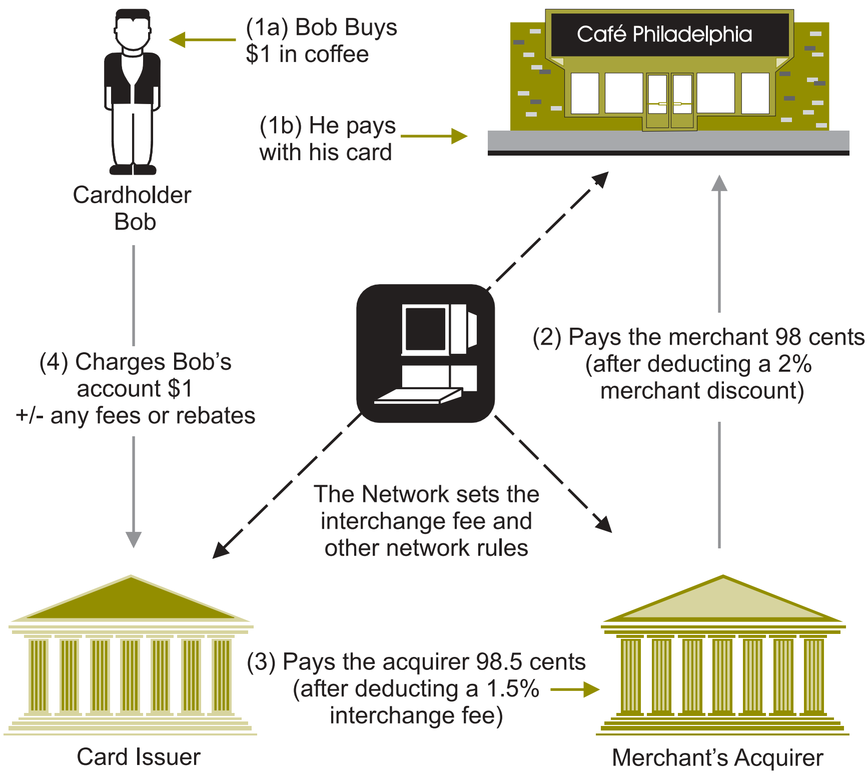
Merchants also have banks, called *acquirers*, that process card payments on their behalf.[4] Merchants pay their acquirer for these services by accepting a *merchant discount* — when a consumer makes a $1 purchase using a payment card, the acquiring bank pays the merchant slightly less than $1 for that transaction (Figure 1).

An *open payment network*, like the bankcard associations Visa and MasterCard and most electronic funds transfer (EFT) networks, allows many banks to participate. The association builds and maintains much of the infrastructure: the lines and switches required to route transaction information between different acquiring and issuing banks. The associations specify that, for each transaction, an *interchange fee* be paid to the bank issuing a card by the bank acting as the acquirer for the merchant.[5] In the U.S., about 1.5 percent of the value of all general-purpose-card transactions flows to issuers in the form of interchange fees — about $23 billion a year.[6]

An interchange fee is one way to ensure that network participants are

FIGURE 1

Flow of Payments in a Stylized Card Network*



(1a) Bob Buys $1 in coffee
Café Philadelphia
(1b) He pays with his card
Cardholder Bob
(4) Charges Bob's account $1 +/- any fees or rebates
(2) Pays the merchant 98 cents (after deducting a 2% merchant discount)
The Network sets the interchange fee and other network rules
(3) Pays the acquirer 98.5 cents (after deducting a 1.5% interchange fee)
Card Issuer
Merchant's Acquirer

* This figure is an illustration; it is not a precise description of any actual network. In particular, it does not reflect the timing of actions required to authorize or settle a transaction.

In the figure, the card issuer makes a payment directly to the merchant's bank. In some networks, this is not done directly, but instead is done via payments to and from the network itself.

The merchant discount and interchange fee are simply illustrative. Other fees (e.g., switch fees) are not included in the figure.

[2] The organization and development of the U.S. payment card industry is described in the book by David Evans and Richard Schmalensee and the book by Lewis Mandell.

[3] In this article, I will focus on general-purpose credit cards, such as Visa or MasterCard, and debit cards. I will not discuss department store cards, oil company cards, or bank cards when they are used at ATMs.

[4] They are called acquirers because they acquire transactions for the network.

[5] In a *closed* network, the card issuer also acts as the acquirer. Such networks carry a merchant discount but not an interchange fee. Examples include American Express and Discover.

[6] This estimate is from *The Nilson Report*, February 2002 (No. 758).

able to recover their costs. But, as we will see, it can also be used to coordinate the activities of banks that issue payment cards and acquiring banks that process transactions on behalf of merchants. Even though consumers do not directly pay the interchange fee, it often affects the cost and benefits of using a payment card. For example, issuing banks that receive higher interchange fees will have an incentive to reduce fees that cardholders pay (annual fees, transaction fees, or interest rates). Or they may offer incentives such as cash back or frequent flyer miles. The interchange fee also affects the acquiring bank because the bank must cover that fee through the discount it charges merchants. That, in turn, influences merchants' willingness to accept cards.

The bankcard association also acts as the rule-making body for the network. In recent years, two of these rules have received a great deal of attention. First, the *honor-all-cards rule* says that merchants wishing to accept a card brand must accept all cards issued under that brand. For example, a merchant who accepts a Visa card issued by ABC Bank must also accept Visa cards issued by XYZ Bank. In addition, merchants must accept all types of a particular brand — from platinum to plain vanilla cards.[7]

Second, the no-surcharge rule says that merchants may not charge customers more for a transaction using one brand of card than for a transaction involving any other brand. Taken together, these rules require that merchants treat all cards issued under a given brand equally and must not favor another card brand by offering its users better prices.

The New Kid on the Block: The Debit Card. Debit cards allow customers to pay for goods and services at the point of sale by authorizing a withdrawal from their checking or savings account. In the U.S., debit transactions at the point of sale only became common in the 1990s, but they have increased extremely rapidly. In the 20 years ending in 2000, consumer purchases made via debit cards rose from essentially zero to account for nearly 12 percent of all noncash consumer transactions. During this same period, the share of these transactions paid via credit card increased from 14 to 21 percent; the share paid by check fell from 86 to 59 percent.[8]

> In the U.S., debit transactions at the point of sale only became common in the 1990s, but they have increased extremely rapidly.

Most ATM cards can be used at the point of sale as debit cards. Such transactions are called *PIN debit* transactions because the cardholder must enter her four-digit PIN to authorize the transaction.[9] Funds are then immediately withdrawn from the associated bank account. The transaction itself is routed through an electronic funds transfer (EFT) network, for example, Star, NYCE, and Pulse.[10] But in order to accept PIN debit transactions, the merchant must first install a PIN pad and have a contract with one or more EFT networks. Today, about 1.3 million different store locations can accept a PIN debit transaction.

Visa and MasterCard offer their own brand of debit card, but they function differently. These cards can typically be used for PIN debit transactions, but they can also authorize transactions with just the signature of the cardholder. A purchase paid for in this manner is often called a signature debit transaction. Unlike a PIN debit transaction, a signature debit transaction does not immediately remove funds from the cardholder's account; it typically takes a day or two for the transaction to clear. This delay creates some credit risk for the issuing bank because the cardholder may have insufficient funds in her account at the time the transaction clears. So, unlike with ATM cards, banks offer signature debit cards only to account holders that meet minimum credit standards.

Another important difference between the two types of debit transactions is that a signature debit transaction can be carried out with the same equipment used to authorize credit card transactions. In fact, under the honor-all-cards rule, stores that accept Visa or MasterCard must also accept the comparable brand of debit card. Currently, 4.9 million store locations in the U.S. accept one or both of these cards, offering a huge merchant base for signature debit cards.[11] Signature debit transactions are routed over the card associations' network, and the card

[7] Bankcard association rules require merchants to accept their brands of debit cards as well.

[8] See the article by Gerdes and Walton. These statistics refer to the number of transactions, not the value of those transactions.

[9] A PIN, or *personal identification number*, is a four-digit number entered on a keypad at an ATM machine or point-of-sale terminal. Many networks require a PIN because, as long as it remains confidential, a PIN can verify that the card is being used by the authorized cardholder.

[10] These networks are also the backbone of the 350,000 ATM machines around the country.

[11] Information is from *The Nilson Report*, June 2002 (No. 765), in a column entitled "Retailer/Bank Card Lawsuit."

issuer receives an interchange fee comparable to the interchange fee on a credit card transaction.[12]

Network Effects and Fixed Costs in Payment Card Networks. Payment networks function differently from most markets, in part because of *network effects*.[13] A payment card is more valuable to consumers when more merchants accept the card. At the same time, merchants are more willing to accept a card if they know many consumers use it. Every consumer who obtains a card and every retailer who accepts a card increase the value of the network to all other cardholders and all other merchants who accept it. Such decisions create externalities that have a number of implications for the evolution and efficiency of payment networks.[14]

First, consumers and merchants are unlikely to take network effects into account unless such effects are reflected in the prices they pay or the benefits they receive.[15] If these effects are ignored, the payment network is likely to be too small or underutilized.

Second, payment networks exhibit increasing returns to scale. If a network is introduced on a small scale, there is inertia — consumers and merchants have little incentive to join. Since such a network would clearly be unprofitable, it would never be launched. But if a network is launched on a large scale, it's possible that many consumers will carry the card and many stores will accept it. If that happens, even more stores are likely to accept the card, which may induce even more people to carry the card and so on.

Third, network effects suggest that there could be significant barriers to entry, and those barriers may permit established payment networks to maintain prices above costs for some time. The reason is that in order to successfully enter the market, a rival network must do so on a large scale. But the market may not be large enough to support more than a few networks at such a scale.

A second factor that distinguishes payment cards from many, but certainly not all, industries is the large fixed cost associated with establishing a viable payment network. Modern payment card networks require large investments in communications and computing facilities in order to make card transactions convenient to customers and merchants and to minimize fraud. The latter is especially important to card issuers because network rules typically promise to pay merchants for fraudulent transactions as long as the network's procedures are followed. If networks did not make this promise to merchants, fewer merchants would be willing to accept payment cards.[16] But card issuers will accept the cost of fraudulent transactions only if the network's antifraud technology is sufficiently effective.[17]

Network effects and fixed costs at the network level may explain why bankcard associations and EFT networks use many of the strategies described earlier. Setting an appropriate interchange fee is one way to ensure that network members take into account network effects, presumably increasing card usage. This in turn reduces the unit cost of card transactions, making payment cards more attractive to use or accept.

Setting the fee at the network level eliminates costs associated with bargaining between individual card issuers and acquirers and uncertainty about the actual costs of a card transaction.[18] Consumers are more likely to use

Modern payment card networks require large investments in communications and computing facilities in order to make card transactions convenient to customers and merchants and to minimize fraud.

[12] A $40 signature debit transaction generates an interchange fee of about 60 cents (1.5 percent) while a comparable PIN debit transaction generates an interchange fee of about 18 cents (0.5 percent). See the August 8, 2002 edition of *ATM and Debit News*.

[13] A nontechnical discussion of this subject can be found in the "Symposium on Network Externalities." For applications of the theory to financial networks, see the article by Nicholas Economides and the one by James McAndrews.

[14] An externality exists when the decisions or activities of one entity affect, positively or negatively, the environment of another.

[15] For example, consumers are sometimes offered cash back or frequent flyer miles as an incentive to use their cards.

[16] An exception to this rule for credit cards is card-not-present transactions, such as Internet or telephone orders, where network rules stipulate that fraud losses are borne by the merchant.

[17] As with network effects, large fixed costs imply economies of scale. This could explain why we did not observe an increase in the number of payment networks in the 1990s (in fact, there was a significant decline), even though technological advances significantly reduced the merchant's cost of accepting a new payment card.

[18] Why doesn't the collective setting of the interchange fee — an obvious example of price fixing among competitors — violate U.S. antitrust law? Federal courts have recognized there are situations where such arrangements may promote rather than retard competition. See the *Broadcast Music, Inc.* and *NaBANCO* cases.

a payment card if they know where it will be accepted and on what terms. The honor-all-cards rule and the no-surcharge rule reduce the uncertainty consumers would otherwise face. This was especially important in the late 1960s and 1970s, when the card associations were trying to build nationwide acceptance of credit cards issued primarily by small banks. But do these mechanisms remain essential after a payment card network becomes well established?

There could be a dark side to all this coordination: These rules might

Legal and Regulatory Challenges to Payment Card Networks

United States. In October 1996, Wal-Mart and other retailers filed an antitrust suit against Visa and MasterCard.[a] This suit later became a class action, representing several million retail locations. In April 2003, the district court ruled on the pre-trial motions, reaching a number of conclusions favorable to the plaintiffs' tying claim.[b] The case was settled shortly thereafter. The bankcard associations agreed to revise their honor-all-cards rules so that merchants can separately decide whether to accept their brands of credit and debit cards, to reduce interchange fees charged on signature debit transactions, and to pay $3 billion in damages over a 10-year period. The reduction in interchange fees in 2003 alone is expected to save merchants $1 billion.[c]

In 1998 the U.S. Department of Justice (DOJ) filed a separate antitrust suit against Visa and MasterCard. Among other things, DOJ objected to the associations' exclusivity rules, which prevent banks that issue Visa or MasterCard credit cards from simultaneously issuing a Discover or American Express card. In October 2001, the trial court invalidated these rules.[d] The case is currently under appeal.

Europe. In July 2002, the Competition Directorate of the European Commission announced a settlement with Visa that addresses multilateral interchange fees levied on certain credit and debit transactions that involve banks in more than one member state.[e] Under the terms of the agreement, Visa pledges to reduce those fees gradually over the next five years and to keep them below a cap that will be calculated each year on the basis of card issuers' costs. Allowable costs include transaction processing, financing the interest-free period enjoyed by cardholders, and certain payment guarantees provided to merchants. Visa also agreed to amend its rules so that its interchange fee can be disclosed to merchants.

In a separate decision published in November 2001, the Commission concluded that Visa's honor-all-cards rule did not restrict competition even when applied to different types of cards (for example, credit and debit) within the same brand (for example, Visa).[f]

Australia. In August 2002, the Reserve Bank of Australia (RBA) announced regulations that apply to domestic credit card transactions using Visa, MasterCard, or Bankcard credit cards.[g] As of January 2003, merchants are permitted to surcharge transactions using these cards. In October 2003, credit card interchange fees will be capped according to a cost-based formula that will be revised every three years.[h] Allowable costs include authorizing and processing transactions, financing the interest-free grace period enjoyed by cardholders, and costs resulting from card fraud and its prevention. The card associations must provide RBA with audited data on these costs each year. RBA also invalidated certain card association rules that it concluded were inhibiting entry by monoline credit card banks and merchant acquirers.

[a] *In re Visa Check/MasterMoney Antitrust Litigation*, N0. 00-7699 (2d Cir 2001).

[b] *In re Visa Check/MasterMoney Antitrust Litigation*, 96-CV-5238 (E.D.N.Y. 2003).

[c] "MasterCard, Visa to Pay $3 Billion to Resolve Card Suit; Will Modify Debit Card Policy, Fees," *BNA Banking Report*, Vol. 80 (May 5, 2003), pp. 739-40.

[d] *U.S. v. Visa U.S.A.*, Inc. 163 F. Supp. 2d. 322 (S.D. NY 2001).

[e] Case No. COMP/29.373 — Visa International-Multilateral Interchange Fees. *Official Journal of the European Community* (July 24, 002).

[f] Case No. COMP/29.373 — Visa International. *Official Journal of the European Community* (November 10, 2001).

[g] Reserve Bank of Australia. "Reform of Credit Card Schemes in Australia IV: Final Reforms and Regulation Impact Statement," August 2002.

[h] RBA is imposing the cap against the volume weighted average interchange fee levied on card transactions rather than specifying caps for different kinds of transactions. RBA expects that once implemented, the caps will reduce average interchange fees about 40 percent.

be used to enhance a dominant network's market power. Such allegations form the basis of an important antitrust case in the U.S., regulation of the payment card industry in Australia, and far-reaching inquiries in Europe (see *Legal and Regulatory Challenges to Payment Card Networks*). In the following sections, I'll examine in greater detail the role of interchange fees, the no-surcharge rule, and the-honor-all-cards rule in consumer payment networks.

THE PROS AND CONS OF INTERCHANGE FEES

An interchange fee can be used to solve a seemingly intractable problem: how to maximize the value of a payment network while ensuring that retailers and banks are able to cover their costs. Economic theory offers some intuition about solving this problem. All other things equal, prices should be set lower for customers who are more price sensitive, that is, more likely to switch to another form of payment in response to a small change in price. Conversely, prices should be set higher for those customers who are not as sensitive to price differences. Economists refer to this strategy as Ramsey pricing, in honor of the mathematician and economist Frank Ramsey.[19] Intuitively, this rule leaves consumers as close as possible to the consumption choices they would make if they were able to purchase goods and services at a price equal to their marginal cost of production — the competitive ideal.

> An interchange fee can be used to solve a seemingly intractable problem: how to maximize the value of a payment network while ensuring that retailers and banks are able to cover their costs.

But open payment networks do not actually control all the prices that consumers and retailers pay for a transaction. Instead, they influence those prices by setting the interchange fee. For example, suppose the network raises the interchange fee so that each card transaction is more profitable for card issuers. Card issuers will seek out more cardholders either by offering them more benefits or by reducing cardholder fees. Merchants will observe more cardholders using the card. But there is a trade-off to raising the interchange fee because it raises costs for the acquiring banks, and at least some of that cost is passed on to merchants. The higher cost of card transactions may cause some merchants to stop accepting the card.

To summarize, economic theory suggests two factors that are likely to influence the size of a *privately optimal* interchange fee, that is, one that maximizes the value of a payment network. The first is the relative size of costs borne by issuing and acquiring banks — banks will not willingly participate if they cannot recover their costs. The second is the degree of price sensitivity exhibited by cardholders on the one hand and merchants on the other. In order to maximize the value of a payment card network, it is necessary to impose more of the costs on those participants who are least likely to stop using or accepting the card.

Networks May Not Choose the Interchange Fee Best for Society. But the best interchange fee for a particular payment network need not be the best from the standpoint of consumers.[20] The economic literature has explored a variety of reasons a privately determined interchange fee may not correspond to the fee that maximizes social welfare, but in this article, we'll focus on just one.[21]

Suppose that a merchant charges the same prices regardless of the manner in which consumers pay. For example, customers who pay with a credit card pay the same price as customers who pay with cash. While different payment instruments mean different costs for the merchant, those costs are not reflected in the prices paid by every customer. Users of the cheaper form of payment bear some of the costs created by customers who use a more expensive form of payment. In economic terms, there is a *cross-subsidy*. Consumers tend to overuse the more expensive form of payment because they

[19] Ramsey solved the following problem: A fixed amount of revenues must be raised by charging prices for goods in excess of their marginal cost. But higher prices will reduce consumption and therefore welfare. Under these circumstances, the best that can be done is to charge higher markups on those goods with less elastic demand curves and lower markups on those goods with more elastic demand curves. See Ramsey's 1927 article.

[20] A payment network that enjoys market power has some freedom to choose the amount of resources to raise through some combination of fees and discounts to merchants and cardholders. When there is more than one payment network, it is not necessarily efficient to encourage the growth of a network if it is at the expense of a less costly one.

[21] See the articles by William Baxter; Dennis Carlton and Alan Frankel; Sujit Chakravorti and Ted To; Sujit Chakravorti and William Emmons; Howard Chang and David Evans; Joshua Gans and Steven King; Jean-Charles Rochet and Jean Tirole; Richard Schmalensee; and Julian Wright.

enjoy all the benefits but do not bear all the costs. At the economywide level, this could mean a payment network will grow too large because purchases outside the network are subsidizing purchases made within the network.

It's possible a network can exploit this cross-subsidy by raising interchange fees while reducing cardholder fees. Merchants would pass on these costs to all their customers, increasing the subsidy noncard users pay to card users. If some of the increased interchange revenues are passed on to cardholders (through lower fees or more perks), this would, in turn, increase the number of cardholders. Merchants may not like this outcome, but they may be reluctant to stop accepting the card if they think cardholders will take their business elsewhere.

The actual outcome depends crucially on how consumers react to changes in prices and the nature of competition among merchants. If customers who do not use the card respond to small price increases by switching to merchants that accept only cheaper cards, or just cash, any cross-subsidy must be small. Thus, an important insight gleaned from theoretical models of payment networks is that the effects of interchange fees depend on the extent of market power enjoyed by retailers.

THE PROS AND CONS OF SURCHARGES

So far, we've assumed that merchants do not set different prices for different card transactions. What happens if payment networks permit merchants to add a fee to transactions when consumers use a more expensive payment method? In principle, merchants could pass on any difference in their cost of using different payment cards to the customers using those cards. This would eliminate any cross-subsidy between customers using different payment methods and encourage consumers to use the most efficient forms of payment. So if we think that a costly payment instrument is being used too much, allowing merchants to surcharge may be a useful remedy.

But permitting surcharges may have a second effect. If merchants are willing to pass on costs in this way, an open payment network cannot use an interchange fee to influence the prices paid by merchants and consumers. To see this, imagine what would happen if the network raised the interchange fee and card issuers passed on the additional revenue to cardholders via lower fees or additional perks. Suppose also that card acquirers pass on the higher interchange fee to retailers by raising the merchant discount. In turn, merchants could increase the surcharge on card transactions, essentially negating the increased benefits provided by card issuers. So with surcharging, raising interchange fees in order to stimulate card use will not be very effective. If network effects are important and cannot be taken into account by some other means, the result could be underutilization of payment cards.

An important insight gleaned from theoretical models of payment networks is that the effects of interchange fees depend on the extent of market power enjoyed by retailers.

But so far we have assumed that given the opportunity, merchants actually would impose surcharges. There is reason to doubt much surcharging would occur. In the U.S., federal law has permitted merchants to offer a discount for cash purchases since 1975. Yet, in 1983, less than 10 percent of retailers offered cash discounts. Most of this activity occurred at gas stations and even this became less common once these stores began to accept bank-issued credit and debit cards.[22] In Sweden and the Netherlands, two countries that banned no-surcharge rules during the 1990s, less than 10 percent of retailers report surcharging their customers and there has been little change in the merchant discount.[23] In practice, then, permitting surcharging may have little effect.

THE PROS AND CONS OF AN HONOR-ALL-CARDS RULE

The question regarding an honor-all-cards rule is not whether networks should be permitted to have such a rule, but rather how broadly it can be applied. Suppose a network issues two types of cards — a credit card and a debit card — under the same brand. Should a merchant be required to accept both types of cards even if it prefers to accept only one type?

This was the central argument in the Wal-Mart antitrust case: The plaintiffs argued that the bankcard associations were using the honor-all-

[22] Statistics on cash discounting are from *Credit Cards in the U.S. Economy*. Evans and Schmalensee document the subsequent decline in cash discounts in their 1999 book. Edmund Kitch argues that regulatory barriers made it relatively costly for merchants to offer cash discounts. Alan Frankel argues that merchants believe any benefit of charging different prices is not worth risking a negative reaction from customers. He also wonders whether consumers react more strongly to fees charged at the point of sale than to fees that appear later on their bank statement.

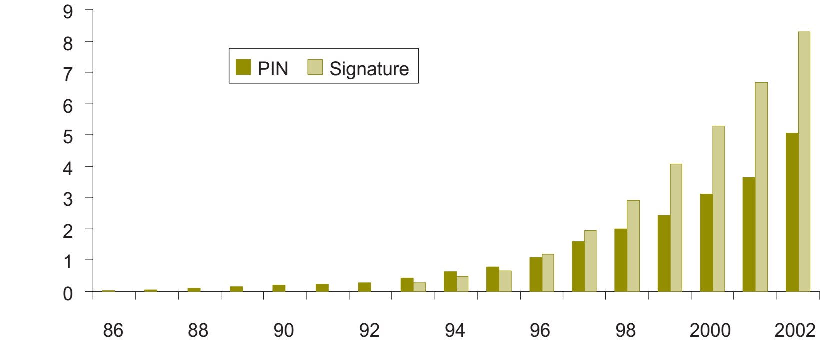
[23] See the November 2001 European Commission decision. In his 2001 report, Michael Katz notes that among retailers in the Netherlands who were aware of the legal change, 20 percent surcharged.

cards rule to impose an illegal tying arrangement, forcing merchants that accept a brand of credit card to accept the same brand of signature debit card.[24] Because card issuers receive more interchange revenue from signature debit transactions, they have an incentive to subsidize consumers' use of these cards. This, in turn, influences consumers' choices about signature vs. PIN debit transactions. Merchants pay these higher fees and pass at least some of the cost on to consumers via higher prices. During the 1990s, the use of both types of debit cards grew immensely (Figure 2). But the share of all debit transactions using a PIN fell from about 60 percent in 1993 to about 38 percent in 2002.

FIGURE 2

PIN Debit vs. Signature Debit Transaction Volume

(billions per year)



Sources: EFT Network Data Book, Debit Card and POS Market Data Book, and Card Industry Directory, various years.

If we view credit and debit cards as distinct products, it seems reasonable that an honor-all-cards rule should be enforced separately for each type of card. In other words, merchants could be allowed to refuse the debit card but still be required to accept all credit cards issued under that brand. Similarly, merchants could separately decide whether to accept a debit card brand but would then be required to accept all debit cards issued under that brand.

The likely result would be that merchants would pay different fees for credit card and signature debit transactions. If signature and PIN debit transactions offer merchants the same benefits, the interchange fee for a signature debit would have to fall in order for these cards to remain competitive with a PIN debit, since the cost of processing a PIN debit is less than the cost of processing a signature debit. Users of signature debit cards would presumably pay higher fees or enjoy fewer benefits, since the use of the cards could not be subsidized as heavily.[25]

Do Credit and Debit Cards Compete in the Same Market?

Applying the honor-all-cards rule to both credit and debit cards of the same brand seems more reasonable if these products actually compete in the same market. But do they? Economists typically define the boundaries of a market based on how consumers respond to price changes. If a change in the price of one good induces consumers to switch to another good, we say these goods are *substitutes.* If only a small change in price is sufficient to cause consumers to switch to the substitute product, we say they compete in the same market.

What can we say about consumer substitution between credit cards and debit cards? There is at least some evidence that consumers do not use credit and debit cards in the same way. For example, consumers are more likely to use debit cards in drug and grocery stores than they are at department stores (Table). In addition, consumers can often initiate a PIN debit transaction larger than their purchase and receive the difference in cash, a feature not available in a credit card transaction. While credit cards provide an explicit line of credit, debit cards do not. Such differences suggest that credit and debit cards are not pure substitutes as a means of payment.

In 2001, a federal court, in a separate antitrust case against the bankcard associations, reached just this conclusion.[26] But in a number of previous antitrust decisions, the courts

[24] In a tying case, the plaintiff tries to prove that the defendant is using the market power it enjoys in one market to extract profits from another, typically more competitive, market.

[25] The Wal-Mart case was settled in April 2003. (See *Legal and Regulatory Challenges to Payment Card Networks.*) Industry analysts predict it will change the debit card market in precisely the way described in the preceding two paragraphs.

[26] This case was initiated by the U.S. Department of Justice in 1998. A similar conclusion was reached in a ruling on pretrial motions in the Wal-Mart case (see *Legal and Regulatory Challenges to Payment Card Networks*).

TABLE

Data on the Use of Debit and Other Forms of Payment (for 1999)

	Stores with PIN Pads (percent)	Percentage of Store Sales Paid via			
		Cash	Check	Credit Card	Debit Card
All Stores	50	35	21	25	8
Discount	43	47	17	27	3
Drug	73	41	17	26	14
Supermarket	100	44	32	11	12
Department Store	20	29	15	26	2
Home Center	7	21	27	26	6
Apparel	38	28	19	32	10

Source: "Survey of Retail Payment Systems," *Chain Store Age* (December 1999)
Note: These statistics are derived from a survey of large retail chains. It is not a representative sample of the retail sector.

have defined the relevant market more broadly to include cash, checks, department store cards, and ATM cards.[27] Even in the 2001 decision, the judge recognized that the emergence of all-in-one cards — a single card that can be used for credit, signature debit, or PIN debit transactions — may increase consumers' willingness to substitute between these different forms of payment.

YESTERDAY, TODAY, AND TOMORROW

At the end of the day, policymakers need to know the answer to the following question: Does the conduct of a payment network benefit or harm consumers? In antitrust cases, judges are often forced to weigh the static costs of certain conduct against any dynamic benefits it may offer. This is not easy to do when it is not clear how a market would have developed in the absence of the conduct under scrutiny.

Suppose we return to the 1980s, before a thriving debit card market developed. How might such a market be developed? The method chosen by Visa and MasterCard was to graft debit cards on to the existing credit card networks. Using their *honor-all-cards* rule, the associations ensured that millions of merchants would accept signature debit cards. Using their *no-surcharge* rule, the associations ensured that these cards would be accepted on terms equal to those of any other debit card. Even with these advantages, signature debit cards were not immediately successful. Their success occurred only after credit cards were commonly accepted in more price-sensitive retail segments, and virtually all merchants were using modern electronic terminals to authorize transactions.[28]

When network effects and dynamic issues are both important, as they appear to be in this industry, policymakers face a difficult problem in deciding what remedies, if any, will benefit consumers in the long run. On the one hand, network rules and pricing strategies may be essential elements in the successful launch of a payment card network and its subsequent expansion. On the other hand, once a payment network is well established, it is possible the same rules can lead to their overutilization and to pricing well in excess of costs. A further complication is that the pricing strategy of existing payment networks affects how and when newer and presumably better forms of payment emerge. New forms of payment must overcome any subsidies consumers receive when using today's payment instruments, and this may require offering subsidies of their own.[29] Policymakers should take all of these factors into account when examining competition among consumer payment networks.

[27] See the NaBANCO decisions.

[28] Both of these developments were promoted by offering interchange fees below the standard rate. For a more detailed description of the evolution of the debit card market in the U.S., see the book by Evans and Schmalensee and the article by Steven Felgran and the one by Felgran and R. Edward Ferguson.

[29] John Caskey and Gordon Sellon argue that the adoption of debit cards in the U.S. was delayed in part by subsidies resulting from the pricing of consumer check transactions. Today, higher interchange fees paid on debit transactions coincide with debit cards' displacement of checks in many consumer transactions.

REFERENCES

Baxter, William F. "Bank Interchange of Transactional Paper: Legal and Economic Perspectives," *Journal of Law and Economics*, 26, 1983, pp. 541-88.

BNA Banking Report, "MasterCard, Visa to Pay $3 Billion to Resolve Card Suit; Will Modify Debit Card Policy," Vol. 80, May 5, 2003, pp. 739-40.

Carlton, Dennis W., and Alan S. Frankel. "The Antitrust Economics of Credit Card Networks," *Antitrust Law Journal*, 63, 1995, pp. 643-68.

Caskey, John P., and Gordon H. Sellon, Jr. "Is the Debit Card Revolution Finally Here?" Federal Reserve Bank of Kansas City *Economic Review*, Fourth Quarter 1994, pp. 79-95.

Chakravorti, Sujit, and William R. Emmons. "Who Pays for Credit Cards?" Federal Reserve Bank of Chicago Public Policy Series, February 2001 (EPS-2001-1).

Chakravorti, Sujit, and Ted To. "A Theory of Credit Cards," mimeo, Federal Reserve Bank of Chicago, 2002.

Chang, Howard, and David S. Evans. "The Competitive Effects of the Collective Setting of Interchange Fees by Payment Card Systems," *Antitrust Bulletin*, 45, 2000, pp. 641-77.

Credit Cards in the U.S. Economy: Their Impact on Costs, Prices, and Retail Sales. Washington: Board of Governors of the Federal Reserve System, 1983.

Debit and Credit Card Schemes in Australia: A Study of Interchange Fees and Access. Reserve Bank of Australia and the Australian Competition and Consumer Commission, 2000.

Economides, Nicholas S. "Network Economics with Application to Finance," in *Financial Markets, Institutions & Instruments*, Vol. 2, 1993, pp. 89-97.

Evans, David S., and Richard Schmalensee. "Some Economic Aspects of Antitrust Analysis in Dynamically Competitive Industries," NBER Working Paper No. 8268 (2001).

Evans, David S., and Richard Schmalensee. *Paying with Plastic: The Digital Revolution in Buying and Borrowing*. Cambridge, MA: MIT Press, 1999.

Evans, David S., and Richard Schmalensee. "Economic Aspects of Payment Card Systems and Antitrust Policy Toward Joint Ventures," *Antitrust Law Journal*, Vol. 63, 1995, pp. 861-901.

Felgran, Steven D. "From ATM to POS Networks: Branching, Access, and Pricing," *New England Economic Review*, May/June 1985, pp. 44-61.

Felgran, Steven D., and R. Edward Ferguson. "The Evolution of Retail EFT Networks," *New England Economic Review*, July/August 1986, pp. 42-56.

Frankel, Alan S. "Monopoly and Competition in the Supply and Exchange of Money," *Antitrust Law Journal*, Vol. 66, 1998.

Gans, Joshua, and Steven P. King. "The Neutrality of Interchange Fees in Payment Systems," mimeo, University of Melbourne, 2001.

Gans, Joshua S., and Steven P. King. "Regulating Interchange Fees in Payment Systems," mimeo, University of Melbourne, 2001.

Gerdes, Geoffrey R., and Jack K. Walton II. "The Use of Checks and Other Noncash Payment Instruments in the United States," *Federal Reserve Bulletin*, August 2002, pp. 360-74.

Katz, Michael L. "Reform of Credit Card Schemes in Australia II: Commissioned Report," Reserve Bank of Australia, August 2001.

Kitch, Edmund W. "The Framing Hypothesis: Is It Supported by Credit Card Issuer Opposition to a Surcharge on a Cash Price?" *Journal of Law, Economics and Organization*, Vol. 6, 1990, pp. 217-33.

Mandell, Lewis. *The Credit Card Industry: A History*. Boston: Twayne Publishers, 1990.

McAndrews, James J. "Network Issues and Payment Systems," Federal Reserve Bank of Philadelphia *Business Review*, November/ December, 1997, pp. 15-25.

Nilson Report, The. February 2002 (no. 758) and June 2002 (No. 768).

Ramsey, Frank. "A Contribution to the Theory of Optimal Taxation," *Economic Journal*, Vol. 31, 1927, pp. 47-61.

Reserve Bank of Australia. "Debit and Credit Card Schemes in Australia: A Study of Interchange Fees and Access," October 2000.

Reserve Bank of Australia. "Reform of Credit Card Schemes in Australia IV: Final Reforms and Regulation Impact Statement," August 2002.

Rochet, Jean-Charles, and Jean Tirole. "Cooperation Among Competitors: The Economics of Payment Card Associations," Centre for Economic Policy Research, Discussion Paper 2101, 1999.

Schmalensee, Richard. "Payment Systems and Interchange Fees," *Journal of Industrial Economics*, Vol. L, 2002, pp. 103-22.

"Study Regarding the Effects of the Abolition of the Non-Discrimination Rule in Sweden," IMA Market Development, AB (February 2000).

"Survey of Retail Payment Systems," *Chain Store Age* (December 1999).

"Symposium on Network Externalities," *Journal of Economic Perspectives*, Vol. 8, 1994, pp. 93-133.

Wright, Julian. "Optimal Card Payment Systems," mimeo, University of Auckland, 2002.

Wright, Julian. "The Determinants of Optimal Interchange Fees in Payment Systems," University of Auckland, Department of Economics Working Paper 220, 2001.

Cases

Broadcast Music, Inc. v. Columbia Broadcasting Co., 441 U.S. 1 (1979)

Case No. COMP/29.373 — Visa International. *Official Journal of the European Community* (November 10, 2001)

Case No. COMP/29.373 — Visa International-Multilateral Interchange Fees. *Official Journal of the European Community* (July 24, 2002)

In re Visa Check/MasterMoney Antitrust Litigation, No. 00-7699 (2d Cir 2001)

In re Visa Check/MasterMoney Antitrust Litigation, 96-CV-5238 (E.D.N.Y. 2003)

National Bancard Corp. (NaBANCO) v. Visa U.S.A., Inc. 596 F. Supp. 1231 (S.D. FL 1984), 779 F2d 592 (11th Cir 1986)

U.S. v. Visa U.S.A., Inc. 163 F. Supp. 2d. 322 (S.D. NY 2001)

Should Philadelphia's Suburbs Help Their Central City?

BY ROBERT P. INMAN

The United States is unique in its commitment to local government as the primary provider of essential public services and in its use of local taxes as the primary means for paying for these services. The Philadelphia metropolitan area is typical of the U.S. pattern. But the city of Philadelphia faces the burdens and responsibilities of all older central cities, including a higher proportion of poor residents than its surrounding suburbs. Such circumstances lead the city to impose higher taxes on city residents, workers, and businesses. Raising revenues through higher taxes, however, becomes self-defeating when tax rates drive people and businesses away. The result is a weaker city and regional economy. How can Philadelphia strengthen its finances? Bob Inman proposes a targeted program of suburban assistance to lower the commuter wage tax and presents evidence that such a program is likely to benefit city and suburban residents alike.

Should the residents of Philadelphia's suburbs — Bucks, Chester, Delaware, and Montgomery counties — contribute to the financing of services provided to city residents and businesses? The United States is unique in its commitment to local government as the primary provider of essential public services and in its use of local taxes as the central means for paying for these services. The Philadelphia metropolitan area is typical of the U.S. pattern. In the five counties that comprise the Pennsylvania portion of the Philadelphia metropolitan area, there are 243 municipal governments and 62 separate school districts servicing a combined population of 3.85 million residents.[1] Local property taxes and, particularly in the case of the Philadelphia region, local resident wage and income taxes are the primary sources of locally raised revenues. In the Delaware Valley, property taxes account for 58 percent and wage/income taxes an additional 28 percent of all locally raised revenues. Locally raised revenues pay for 77 percent of local government and school district spending in the five-county Philadelphia area. Fiscal transfers from the state cover most of the remaining 23 percent.

Bob Inman is Miller-Sherrerd Professor of Finance and Economics, Wharton School, University of Pennsylvania. When he wrote this article, he was a visiting scholar in the Research Department of the Philadelphia Fed.

While only one of many local governments in the metropolitan area, the city of Philadelphia is arguably the region's economic, cultural, and entertainment center. The city has 34 percent of the five-county region's jobs. There are 12 Fortune 500 corporate headquarters in the five-county area, and eight of those headquarters are located in Philadelphia. Four of the nation's 100 largest law firms have their home offices in Philadelphia. Philadelphia's four medical schools are national leaders in patient care and medical research. Together, the four schools currently receive more than $550 million a year in National Institutes of Health funds for faculty research. Higher education is a major industry for the region, and 46 percent of the

[1] There are also four New Jersey counties included in the official definition of the Philadelphia metropolitan statistical area (MSA). For the reasons noted in footnote 9, these counties are not included in this paper's policy analysis.

region's college and graduate school enrollees attend Philadelphia universities. The Philadelphia Orchestra, the Curtis Institute of Music, the Philadelphia Museum of Art, the Franklin Institute, and the Philadelphia Zoo are world-recognized centers of arts and science education. Philadelphia has the area's four major league sports franchises. There are 13 professional theaters providing full seasons, two professional dance companies, and nine music venues featuring artists from major record labels. Of the region's 318 restaurants rated excellent by the 2002 *Zagat's Guide*, 220 are in Philadelphia.

Though it is the economic and entertainment center of the region, Philadelphia also faces the burdens and responsibilities of all older central cities. Philadelphia is home to most of the region's poor and elderly households. While the city has 39 percent of the region's population, it has 70 percent of the region's poor. Philadelphia residents also face significantly higher rates of crime. In 1998, the rate of violent crime was 1465 per 100,000 residents in Philadelphia, yet only 286 per 100,000 residents in the suburbs; the rate of property crime was 5855 per 100,000 in Philadelphia compared with 2503 per 100,000 in the suburbs. These higher service burdens from poverty and crime necessarily translate into higher city tax rates. The overall tax burden for a typical homeowner in Philadelphia is 14.4 percent of family income but only 9.5 percent for an identical family living and working in the suburbs. A similar differential tax burden holds for a typical city firm. Leaving the city for a suburban location will lower a firm's effective state and local tax burden on profits from 16.5 percent to 13.2 percent.[2]

The unique fiscal burdens that Philadelphia faces contribute importantly to these tax differentials. The danger is that these added burdens — and resulting higher tax rates — will undermine the city's economic, cultural, and entertainment advantages to the detriment of all residents in the five-county region. Large cities often have significant cost advantages, known as agglomeration economies, in producing and providing goods and services. These agglomeration economies arise when firms, retail stores, or cultural activities are concentrated in common and usually small geographic areas within the city. High tax rates and low quality public services, however, may drive firms and middle and upper income households from the city. As firms and families exit, city agglomeration economies are lost. The loss of agglomeration economies leads to higher prices for city-produced goods and services. Population and income grow more slowly or decline, and house values in the city and suburbs fall. In the end, the region as a whole loses, not just the central city.

Though it is the economic and entertainment center of the region, Philadelphia also faces the burdens and responsibilities of all older central cities.

The solution is to strengthen city finances. To this end, suburban residents might wish to make a contribution, most realistically done through adjustments in state assistance to local governments. *If* done correctly, suburban fiscal assistance to the central city can be a high-return investment in suburban jobs, growth, and house values.

Here I offer one proposal for such assistance, a targeted program of suburban assistance to lower the commuter wage tax. With this reform, city and suburban homeowners both gain; regional home values are predicted to rise by $2.1 billion, or about $2250 per family.

THE ECONOMIC CONNECTION BETWEEN CITY AND SUBURBS

There are two potential links from the city economy to suburban residents' economic welfare: jobs and wage income for suburban commuters and the market price for city-produced goods and services purchased by suburban firms and residents. For a typical suburban resident, the primary economic advantage of a strong city economy lies in the ability of city firms to provide goods and services at prices lower than (or, equivalently, at a quality higher than) what might be available from suburban firms or from firms outside the metropolitan region.[3] The central city's economic advantage can arise from either of two sources: natural advantages because of the city's proximity to an important production input such as power or raw materials (e.g., Pittsburgh's history as a steel production center), or agglomeration advantages facilitated by the density of firms and households within the city. For U.S. cities today, the likely source of any advantage is agglomeration economies.

Agglomeration economies benefit producers and consumers of city-produced goods and services. A high

[2] The estimates of local tax burdens are available from my biennial report on local taxation, "Local Taxes and the Economic Future of Philadelphia: 2002 Report."

[3] See the article by Richard Voith for more about the advantages commuter suburbanites derive from a strong city economy, with particular reference to Philadelphia commuters.

density of firms within the same industry — called Marshallian agglomeration in honor of Alfred Marshall's initial analysis — leads to lower shipping costs for firms' inputs when there are economies of scale in transportation (e.g., coal and iron ore to the steel mills of Pittsburgh and Gary). The density of firms may also lower labor costs in industries when laid-off workers from a declining firm are quickly hired by an expanding firm. This will be the case in industries where brand loyalty is weak and current fads define consumer demands, for example, the fashion industry in New York, the entertainment industry in Los Angeles, and the "dot.com" industries of Silicon Valley. Having many firms in the same local labor market reduces the unemployment risk to workers with unique talents and therefore allows all firms in the market to pay a lower wage. For much the same reason, a high density of firms in the same industry may also encourage supplier innovation and specialization, again lowering firms' production costs. Furthermore, low-cost production technologies are likely to be more quickly copied when firms and workers are in close proximity. These idea "spillovers" may occur within or across industries, an advantage called Jacobian agglomeration for Jane Jacobs' insightful analysis of growing city economies. Finally, a high density of households and firms gives rise to agglomeration advantages in retailing and consumer services, for example, dining, specialty shopping, and entertainment. In Philadelphia, South Street, Rittenhouse Square, and the Avenue of the Arts are lively examples.

A growing body of economic research has demonstrated the presence and importance of agglomeration economies in regional economies. This research has found that concentration of industry employment has a statistically significant and quantitatively important effect on plant productivity. For example, Mark Beardsell and Vernon Henderson found that doubling the number of computer firms in a given location increases the productivity of those firms as much as 17 percent; little wonder, then, that the Silicon Valley has become the world's leader for computer industry production and innovation. Though not as dramatic, the findings of Antonio Ciccone and Robert Hall showed that overall employment concentration also improves worker productivity; doubling county-level employment density within a state improves all worker productivity in that state by 6 percent. Ciccone has found a similar effect of employment concentration on worker productivity in European firms.[4] Finally, Stuart Rosenthal and William Strange find that the benefits of employment density occur within small geographical areas and are typically exhausted beyond a distance of five miles. The Rosenthal-Strange results suggest, importantly, that the spatial reach of agglomeration advantages will typically be confined within a political jurisdiction. For historical reasons, that political jurisdiction is most often the region's central city.

The gain to city and suburban residents of living in or near a productive central city comes largely from their ability to buy city-produced goods and services at comparatively low prices. This advantage is larger the more goods and services suburban households and firms buy from city producers and the greater are the cost and shipping advantages those city producers have over their closest competitors. For suburban residents in the Philadelphia metropolitan area, it is cheaper to buy expert legal advice, accounting services, life-saving medical care, or first-run professional entertainment from Philadelphia businesses and venues than from those in New York, Baltimore, or Washington, D.C. James Rauch has shown that the ultimate beneficiaries of access to low-cost goods and services due to agglomeration economies are the region's workers and home owners. Wages and house prices are higher in economically more efficient regions.

The gain to city and suburban residents of living in or near a productive central city comes largely from their ability to buy city-produced goods and services at comparatively low prices.

Low quality public services or high city taxes are one important factor that might undo this economic advantage, however. On this point the evidence is clear. My research for Philadelphia —now confirmed in companion studies for Houston, Minneapolis, and New York City — shows unequivocally that high city taxes unmatched by compensating service benefits will drive middle- and upper-income taxpayers and businesses from

[4] In addition, there is preliminary evidence that an increased density of firms not only increases the initial equilibrium level of regional production and incomes but it may also stimulate higher economic growth as well, primarily through the sharing of ideas and innovation. See the article by Edward Glaeser, Hedi Kallal, Jose Scheinkman, and Andrei Shleifer and the one by Vernon Henderson, Ari Kuncoro, and Matt Turner. In his *Business Review* article, Gerald Carlino has provided a valuable survey of the theory and evidence on agglomeration economies and the economic performance of cities and regions.

the city, first to the surrounding suburbs, but just as likely, to other regions of the country as well.[5] For example, I estimate that Philadelphia lost 207,000 jobs over the past 30 years solely because of increases in the city's wage tax rate. Most damaging for the location of businesses in Philadelphia is the city's nonresident portion of the wage tax, a tax whose burden is largely shifted back onto city businesses as workers with suburban job alternatives require a wage increase to compensate for the city's tax. The recent wage tax cuts proposed by former Philadelphia mayor Ed Rendell and the current mayor, John Street, are estimated to have restored approximately 12,000 of the jobs previously lost. Similar adverse effects of high taxes on city property values and city jobs have shown up in Houston, Minneapolis, and New York City.

The loss of city jobs due to high city taxes will mean a reduction in the city's production advantage because of lost agglomeration economies. In our 2002 study, Andrew Haughwout and I tested empirically the argument that weak public finances in the central city can undermine private-sector economic performance, for both city and suburban residents. Our study seeks to explain the growth in city and suburban incomes, populations, and house values over the decade 1980 to 1990 for the 195 largest metropolitan areas in the United States. Weak city finances are measured in this study by four separate indicators of budgetary pressure: the share of city spending paid for through taxation of businesses and middle-class residents; whether the city is required by state law to bargain with its public employee unions; whether the city lacks mayoral veto control over city budgets; and the fiscal burden of city poverty (Table 1).

Columns (1) and (2) of Table 1 show what happened to house values in the average U.S. central city and its surrounding suburbs because of weak city finances during the 1980s, the most recent decade for which we have complete information.[6] On average, the share of city taxes borne by middle and upper income residents and firms rose 1 percent over the decade (Change in SHARE). The consequence was to depress house values in the average city by $3638, or about 7 percent of the initial 1980 value. Importantly, the house values in the average suburb fell too, by $2468, or about 4 percent of their initial 1980 values. (See *Understanding the Economics Behind Table 1*.)

Cities that are required by state law to hire their labor services exclusively from public employee unions (BARGAIN) typically face higher labor costs, and this raises city taxes. Again, higher taxes without compensating service benefits drive households and firms from the city, and city and suburban house values decline from what they might have been had the city retained the right to hire nonunion public employees — that is, to contract out. The average city house loses $5358, or 11 percent of its initial 1980 value, by being in a strong union city. Suburban house values fall too, now by $4047, or 8 percent of their 1980 values.[7]

The same adverse effects on city and suburban house values, though not as large, are seen when city budgets are decided by a majority of ward-elected city council members, unchecked by a constitutionally strong mayor with veto powers (GOVERNANCE). The budgets of such governments typically favor neighborhood services, with tax burdens allocated toward business. We should expect to see an exit of businesses from the city, lost agglomeration economies, higher prices for city goods, and for this reason, lower city and suburban house values. This is what we find. City house values are lower in weak governance cities by an average of $1948 (4 percent of 1980 values) while house values in suburbs surrounding a weak governance city fall an average of $3035, or 6 percent of 1980 values (see Table 1). Interestingly, in cities with weak fiscal governance, we see suburban house values falling more than city house values, but this too makes sense if city

[5] See the article by Andrew Haughwout, Robert Inman, Steven Craig, and Thomas Luce. The conclusion that high taxes and low services depress local economies is now a well-established fact more generally, and Timothy Bartik provides an excellent review of this research.

[6] Our full study reports the effects of weak city finances on city and suburban incomes and population as well; see Haughwout and Inman (2002). Elsewhere we have shown that changes in house values are the best single predictor of changes in the economic welfare of residents in a metropolitan area, so I shall focus on these results here; see also the 2001 article by Haughwout and Inman.

[7] While the typical city or suburban resident loses by living in or near a strong union city, unionized city workers are net winners. Although the value of a unionized city employee's house falls like the house values of all residents, the worker is compensated by his or her gain in personal income. Estimates of the premium unionized city employees earn above what they might earn in their next best private-sector job are typically 3 to 8 percent of the worker's wage; see the article by Richard Freeman and Robert Valletta. For a unionized city employee earning $30,000 a year, this is an annual wage premium of $900 to $2400 per year. Unionization reduces house values by $5358 in the city and $4047 in the suburbs, implying an equivalent annual loss in value of $270 a year for a city house and $200 per year for a suburban house when interest rates are 5 percent. Conservatively, then, the typical unionized city worker in a strong union city gains $900 to $2400 in wages each year and loses from $200 to $270 a year in declining house value. On balance, city employees gain when working in cities with strong public-sector unions.

TABLE 1

City Finances and Metropolitan Area Home Values

	Estimates from a regression model based on data from 195 U.S. metro areas		Estimates from a simulation model of the Philadelphia area economy	
	Estimated Change in Average U.S. City Home Value (1)	Estimated Change in Average U.S. Surburban Home Value (2)	Estimated Change in Average Philadelphia City Home Value (3)	Estimated Change in Average Philadelphia Suburban Home Value (4)
Change in SHARE	-$3638 (974)	-$2468 (873)	-$265	-$1224
BARGAIN	-$5358 (1739)	-$4047 (1563)	-$7184	-$5902
GOVERNANCE	-$1948 (1052)	-$3035 (946)	—	—
Change in POVERTY	-$12,345 (2460)	-$6696 (2212)	-$410	-$15

Columns 1 and 2: Source: Table 6, Haughwout and Inman (2002). The standard error of the estimated change in the city median house value is presented in parentheses below each estimate. Columns 3 and 4: Estimates of changes in the Philadelphia and suburban median home values are computed using an equilibrium political economy model of the Philadelphia MSA, calibrated to match the Philadelphia MSA for the decade 1980-90. The simulation model is described in an appendix to this article. Estimated changes in city and suburban home values are computed for each of the following four changes in the underlying structure of Philadelphia's public finances: Change in SHARE, allowing for the increase in the middle and upper income families' share of city taxes for 1980-90 because of the change in the percent of Philadelphia's population who are poor or over the age of 65; BARGAIN allowing for a 20 percent increase in the real cost of city services for 1980-90 because of the 32.9 percent increase in the real cost of city workers' compensation over the decade; the constitutional form of city GOVERNANCE remained constant over the decade as the city retained its strong mayor form of government and thus there is no impact on city budgetary costs nor city and suburban home values; Change in POVERTY, allowing for the decline in the city's rate of poverty from 20.6 percent to 20.3 percent and the mandated increase in the city's real (inflation-adjusted) contribution for services for low-income households of approximately $200 per poor household.

residents gain at the expense of city businesses. Unfortunately, what city residents seem to gain from their budget is more than offset by what they lose from the decline in the city's economy.

The most damaging change for our nation's largest cities during the 1980s came from the growth in the rate of urban poverty (Change in POVERTY) and associated increases in city spending and taxes. If the city's rate of poverty increased — and this was the case for most U.S. cities — the resulting added tax burden must be spread over relatively fewer middle-class and

UNDERSTANDING THE ECONOMICS BEHIND TABLE 1

The estimated effects of weak city finances reported in Table 1, columns (1) and (2), are statistical estimates of the effects of weak city finances on city and suburban home values for a sample of the largest 195 U.S. metropolitan areas for the decade 1980-1990. The estimates in Table 1, columns (3) and (4), are estimates for the effects of changes in the same city fiscal variables on Philadelphia city and area suburban home values for the decade 1980-1990 derived from an economic model of the metropolitan area economy. An Appendix to this article provides a brief summary, and Haughwout and Inman's 2002 article provides the technical details. An important question to ask of any statistical estimation or model prediction is: *Do the numbers make economic sense?*

These do. Here is a working example using typical family incomes, consumption, and interest rates for the 1980s. Start with a suburban family that, over the 1980s, spent $50,000 per year, of which $10,000 is allocated to city-produced goods and services. Remember, these goods do not have to be consumed in the city, but simply made there or processed and shipped by city businesses. If a higher city tax share leads to the exit of city businesses and lost agglomeration economies, how much would prices of city-produced goods have to increase to justify a $2500 decrease in suburban house values (Column 2, Table 1)? Prices of city-produced goods would have to rise only 1.25 percent. Suburban residents would pay $125 more per year for their city-produced goods, and they would lose this $125 every year. An annual loss of $125 is economically equivalent to the estimated loss in house value of $2500 when the real interest rate is 5 percent.

What about city residents? They suffer the same losses from high prices for city goods, so their house values should also fall by $2500 because of lost agglomeration. But as seen in the first entry of column 1 in Table 1, the estimated average decline in city house values is $3600. The additional $1100 loss in value must come from the direct adverse effects of higher city tax shares. In our sample, on average, the city budget is $6000 per family and the typical middle-income resident's share of city taxes is about 0.50. So an increase of 0.01 in the tax share means an increase of about $60 each year in tax payments ((0.51-0.50) x $6000 = $60).

But again this is an annual loss that will be economically equivalent to a one-time value loss of $1200 when real interest rates are 5 percent. City residents lose $2500 because of lost agglomeration and an additional $1200 because of higher taxes. In this example, the total loss for a typical city house will then be $3700, again very close to the estimate in Table 1. Similar calculations can be made for all the numbers in Table 1; they are all plausible.

business taxpayers. The rise in city taxes because of the average increase in the rate of city poverty, which was 3 percent, led to an average loss in city house values over the decade of $12,345, a decline of more than 25 percent.

Rising city poverty, rising city taxes, and a shrinking city economy affect the suburbs, too. We estimate the average decline in suburban house values over the decade from the growth in city poverty equaled $6696, or 13 percent of the original 1980 suburban house value. When central cities become poorer or the fiscal burden from each poor household increases, city residents lose, but importantly, so too do residents in the city's surrounding suburbs. The path from city poverty to suburban house values may be roundabout — from higher city poverty and city taxes to a weaker city, then regional, economy — but the impact is significant nonetheless.

Finally, each of the four sources of weak city finances is largely outside the direct control of any city's mayor. To be sure, poor fiscal management by a city's mayor or elected council will also lead to higher city taxes and lower city and suburban property values, but that is not what we are measuring in Table 1. Tax laws, the rules of labor bargaining, the city charter, federal and state mandates, the rate of city poverty — these are the determinants of weak city finances as measured here, and each is given to, not chosen by, the mayor. If a mayor is dealt weak fiscal institutions, it should be no surprise that the city and its region lose economically over time.

THE PHILADELPHIA CONNECTION

The estimated effects of weak city finances on city and suburban house values reported in columns (1) and (2) of Table 1 are for a national average city and its suburbs. While Philadelphia is included in the Haughwout-Inman national study, those results cannot be applied directly to Philadelphia. Perhaps Philadelphia was one of the lucky cities dealt a winning fiscal hand. Unfortunately, my estimates from an economic model of the Philadelphia metropolitan economy show this is not the case.[8] Philadelphia faced many of the same fiscal difficulties during the 1980s as our national sample. The results in columns (3) and (4) of Table 1 show that while the losses were not as large as those felt nationally, Philadelphia and its suburbs suffered significant economic losses during the 1980s from weakened city finances.[9] (See An *Economic Model of Philadelphia and Its Suburbs*.)

First, as was true nationally, the share of Philadelphia taxes borne by its middle- and upper- income households (Change in SHARE) rose 1 percentage point, from 50 percent in 1979 to 51 percent by 1989. The causes in Philadelphia were the aging of the city's population and the city's continued loss of manufacturing jobs. The rise in middle-class tax burdens led to a predicted fall in city and suburban house values (Table 1, columns 3 and 4).

Second, Philadelphia's public employees enjoy an exclusive right to bargain with the city. Thus, Philadelphia qualified as a strong union city (BARGAIN) in the Haughwout-Inman national study reported in columns (1) and (2) of Table 1. Each strong union city in that study had its own bargaining experience with unions, but on average, those unions increased labor costs and city taxes and depressed city and suburban home values. This appears to have been the case in Philadelphia as well. Over the decade, Philadelphia's public-employee unions were able to negotiate labor contracts that increased real — that is, inflation-adjusted — compensation for city employees 33 percent, equal to an annual real rate of growth of 2.89 percent in city workers' pay. This increase in Philadelphia's real costs of public employees' labor services led to higher city taxes and, as was true for the national sample of strong union cities, significantly lower city and suburban house values. Our estimates of the effect that strong public-employee unions have on Philadelphia city and suburban house values appear in columns (3) and (4) of Table 1.

The city's adverse fiscal changes over the 1980s are estimated to have reduced the value of a typical city and suburban house each by about $8000.

Third, there was no change in city governance (GOVERNANCE) over the decade; Philadelphia had, and continues to have, the strong mayor form of city government. This row in Table 1, therefore, shows no changes.

Finally, while the burden of city poverty on Philadelphia's budget is high and has an important negative effect on the performance of the regional economy, as we will see below, the 1980s did not add significantly to that burden. Thus, the net economic effects of the change in the rate of poverty (Change in POVERTY) were small as well. The share of Philadelphians living in poverty fell slightly over the decade from 0.206 to 0.203. There was, however, a small offsetting increase in the city's cost of serving that population.[10] Overall, poverty's fiscal burden on the city rose only slightly, and the estimated additional damage done to

[8] The estimates in columns 3 and 4 of Table 1 for the Philadelphia economy are computed using a general equilibrium simulation model for the Philadelphia metropolitan area, calibrated to match the economic and political structure of the five Pennsylvania counties in the Philadelphia MSA for the decade 1980 to 1990; see the Appendix. Ideally, I would have replicated the statistical analysis of the national city sample for a sample comprising only Philadelphia and its suburbs, but unfortunately, this is not possible because the required number of years of suburban data are not available. In the simulation model, the fundamental economic relationship that determines the results in columns 3 and 4 of Table 1 is the efficiency advantage of city agglomeration economies; the stronger these economies are, the larger will be the adverse effects of weak city public finances. For this analysis, I select a very conservative elasticity of Philadelphia city output with respect to city firm density of only 0.01; the national average elasticity of worker productivity with respect to firm density is 0.06, as estimated by Ciccone and Hall.

[9] The four New Jersey counties of the Philadelphia MSA are not included in the simulation analysis for two reasons. First, New Jersey now rebates the Philadelphia commuter tax for New Jersey suburban residents. This different treatment of an important city tax requires separate analyses for the Pennsylvania and New Jersey suburbs of Philadelphia. Second, since our focus is on the economic returns to reforming city and suburban financing of city services, I have chosen what seemed to be the most politically realistic group of suburban counties to be included in any such reforms. Those counties are in Pennsylvania alone.

city and suburban house values is barely noticeable.

All together, the city's adverse fiscal changes over the 1980s are estimated to have reduced the value of a typical city and suburban house each by about $8000, a loss in value of 9 percent for Philadelphia home owners and 6 percent for suburban home owners.[11] Far and away the most important cause of these economic losses for our region was the increase in city taxes required to fund the significant growth in the real compensation of the city's public employees.

INVESTING IN STRONGER CITY FINANCES: A STRATEGY FOR GROWTH IN THE PHILADELPHIA REGION

Both in the nation and in Philadelphia, weak city finances lead to the exit of mobile city firms and households, a less efficient city economy, and lower incomes and house values for the region as a whole. Strong city finances protect a city's economic efficiency and a region's income and wealth. What might we do to strengthen Philadelphia's city finances, and what will be the gains to city and suburban residents from such a strategy?

Three Possibilities. The analysis in Table 1 identifies three possible directions in which the current structure of Philadelphia's finances might be improved. First, reduce the city's relative tax burden on mobile middle-class households and city firms. Second, control the ability of the city's public employee unions to win favorable compensation packages with greater-than-inflation increases. Third, reduce the city's fiscal obligation for services to lower income households.

On its own, Philadelphia has already made significant progress on two of these three fronts. First, from 1995 to today, the city has lowered its wage-tax rates 9.27 percent and its gross-receipts tax rates 29.3 percent. Increases in the rates of these two taxes over the past 30 years have caused significant damage to the city's economy; so lowering these rates is an important step toward restoring our fiscal competitiveness.[12] Second, since 1992, and in sharp contrast to the 1980s, Philadelphia's compensation per public employee has *declined*, in real dollar terms, at an annual rate of 0.60 percent. The city's improved labor compensation policy has allowed balanced-budget tax reductions, significantly improving the city's ability to attract firms and households. The one important fiscal weakness that has not yet been addressed is the city's continuing high budgetary obligation for support of its poor population.

The direct tax costs needed to fund poverty-related county spending in Philadelphia and in each of the four surrounding Pennsylvania counties are reported in Table 2. Poverty spending in Philadelphia is greater than comparable spending in all suburban counties combined (Table 2). Further, the direct tax burden of poverty spending as a percent of county residents' income is roughly four to seven times higher in Philadelphia than in the suburban counties. While the residents of Bucks, Chester, Delaware, and Montgomery counties pay only 0.17 percent to 0.38 percent of their income in taxes to fund poverty services, Philadelphia residents pay taxes equal to 1.4 percent of their income to fund city-provided poverty services.

The root causes of these spending and tax disparities are, first, the geographical concentration of the region's poor and low-income elderly households within the city, a concentration due in large measure to the availability of older, lower cost housing within the city,[13] and second, the state of Pennsylvania's decision to make counties the primary providers and administrators of poverty-related services. Seventy percent of the five-county region's poor live in Philadelphia, and because the city is also legally a county, Philadelphia must assume primary fiscal responsibility for the unreimbursed portion of the services provided to those families. Philadelphia spends more per taxpayer for poverty services than the suburban counties, not because poverty spending is a successful election strategy or the city's middle class is particularly generous, but because, as the region's oldest and largest city, it has more poor families as residents and because it is a city-county,

Poverty spending in Philadelphia is greater than comparable spending in all suburban counties combined.

[10] See the article by Anita Summers and Lara Jakubowski.

[11] The decline in house values from the combined fiscal changes will not equal the sum of the three isolated changes because the exit of households and firms has accelerating effects in the presence of agglomeration economies. Large adverse fiscal changes will be proportionally more harmful than small changes.

[12] For the most recent analysis of the effects of city taxation on city business, see the article by Haughwout, Inman, Craig, and Luce.

[13] See the article by Edward Glaeser and Joseph Gyourko.

TABLE 2

Direct Tax Cost of Poverty Spending to County Governments in the Philadelphia Region: FY2002

(Millions of Dollars)

County (County % Poor)	Bucks (4.92%) (1)	Chester (4.88%) (2)	Delaware (8.31%) (3)	Montgomery (4.55%) (4)	Philadelphia (19.74%) (5)	Region (11.93%) (6)
PUBLIC HEALTH	$3.816m	$12.951m	$0m	$1.387m[+]	$76.203m	$94.357m
HUMAN SERVICES	$5.198m	$8.169m	$4.992m	$53.069m[+]	$67.993m	$139.421m
CORRECTIONS	$39.485m	$20.828m	$18.634m	$33.102m	$183.120m	$295.169m
EMERGENCY SERVICES	$6.658m	$1.545m	*	*	$15.564m	$23.767m
TOTAL ($ per Non-Poor) (% of Income)	$55.157m ($97.07) (0.34%)	$43.493m ($105.47) (0.32%)	$23.626m ($46.78) (0.17%)	$87.558m ($122.33) (0.38%)	$342.880m ($281.52) (1.4%)	$552.714m ($163.03) (0.60%)

The *direct tax cost to county residents of poverty spending* is defined as county poverty spending minus state and federal grants, departmental earnings, and fees paid to the county for poverty services, all reported in millions of dollars. In Delaware County, for example, all of county spending in FY 2002 for public health was supported by nontax dollars. The total dollar tax burden per nonpoor household ($ per Non-Poor) is calculated as the total poverty spending divided by the population of the county not below the poverty threshold. The percent of county income required to support the county's tax cost of poverty (% of Income) is computed as the total direct tax cost divided by total county residential income.

[+]Montgomery County classifies $37.838 million for geriatric centers as spending within human services; other counties classify such services as part of the public health budget.

* In Delaware and Montgomery counties emergency services for low-income households have been classified as part of the human services budget.

state law demands it.[14]

Regionalization. In this regard, Philadelphia stands in sharp contrast to Pittsburgh. Pittsburgh too is an older city, and the current percent of Pittsburgh's residents who fall below the poverty threshold (19 percent) is almost identical to Philadelphia's. But Pittsburgh's taxpayers share the burden of financing services for their city's poor with the suburban residents of Allegheny County. As a consequence, Pittsburgh city residents face the same tax burden on income as their suburban counterparts, only 0.23 percent.[15] This rate is significantly below the 1.4 percent burden on income now paid by Philadelphia city residents. From the perspective of regional economic growth and welfare, the Pittsburgh metropolitan area has the financing right. Large disparities in the fiscal costs of regional poverty between local jurisdictions discourage firms and households from moving to high poverty locations. If these locations are also the region's important centers of agglomeration economies — as is likely the case in Philadelphia — the firms and households that create those economies leave and economic inefficiency results. The whole region loses.

[14] In a recent study of poverty spending by Philadelphia, I showed that the trend in city spending is unrelated to who is mayor or to the racial and ethnic composition of city council. The main determinant of the city's poverty spending is the performance of the city and national economies. That is, when economic performance improves and there are fewer families in poverty, poverty spending falls.

[15] Allegheny County's direct tax costs for the poverty-related services totaled $64.867 million in fiscal year 2002. The tax burden on a county resident not classified as poor equaled $62.48 per resident. This burden as a percent of county residential income equaled 0.23 percent of county income.

One solution — the Pittsburgh solution — is to regionalize the financing of poverty. The Pittsburgh area achieved this efficient structure for poverty financing by the luck of history. Pittsburgh's historic boundaries define a geographically small city within a geographically large county. The Philadelphia region has not been so lucky. To solve its financing inefficiency, the region must fashion a clear policy for sharing the five counties' cost of regional poverty. If the sharing is done correctly, however, everyone — city and suburban residents alike — can benefit. The best policy will entail a de facto transfer of approximately $191 million a year in poverty relief from the suburbs to the city, with the transfer tied to a required proportional reduction in the city's nonresident wage tax. If implemented, this policy is estimated to increase the combined economic wealth of city and suburban home owners by $2.1 billion, an average of about $2250 per family.

Here is how a policy to transfer funds and reduce the commuter tax rate might work. The budget data in Table 2 allow us to calculate the required suburb-to-city transfer needed to ensure uniform regional financing of regional poverty. To meet the total regional tax burden from poverty of $552.7 million for fiscal year 2002, we need a uniform regional income tax rate of 0.60 percent. Suburban residents have already made a contribution toward their regional share, but in all cases, it is below the uniform regional rate. For example, the taxpayers of Bucks County have already contributed 0.34 percent of county income (Table 2) toward the target contribution of 0.60 percent; so the reform policy would ask those residents to pay an additional 0.26 percent, or in total about $42.622 million, toward the reform policy. Similar calculations can be completed for each of the other three suburban counties.

For fiscal year 2002, the total reform contributions from the four suburban counties would equal $191 million, or about $220 per suburban family.[16] This total would then be paid to the city to lower its poverty-related tax burden. Importantly, no money need actually change hands among the five counties. Since each suburban county receives from the state poverty-related grants and reimbursement revenues greater than its required contribution for uniform regional poverty financing, the state can implement the regional policy by reallocating a portion of suburban grants to Philadelphia. There is no need for regional taxation or regional government to implement poverty-financing reform.

Transfer of Money. How the money is given to Philadelphia matters, however. Table 3 estimates what might happen to the values of typical city and suburban houses and to total house values regionwide when the reform transfer is given to the city in one of three ways. For purposes of comparison, Table 3 also reports census year 2000 house values for the current "No Reform: Status Quo" policy. Reform Policy 1 gives the poverty-relief funding to the city with "no strings attached" — that is, the city is free to allocate the funds any way it wishes.

Reform Policy 2 requires relief funding to be allocated to a uniform percentage reduction in the city's wage tax rates for residents and nonresidents. The proposed level of city poverty relief will permit a 10 percent reduction in each of the two wage-tax rates.

Finally, Reform Policy 3 requires all of city poverty relief to be allocated to reducing the wage-tax rate for nonresidents. This strategy will be particularly valuable to city businesses, since they bear a large share of the burden of the nonresident wage tax as higher labor costs. Reform Policy 3 is likely to be suburban residents' favorite option too, since they benefit most from the larger and more productive city economy that this strategy encourages. Under Reform Policy 3, the nonresident wage-tax rate can be reduced 22 percent.

Reform Policy 1. Under the "no strings attached" policy, the city is free to spend its poverty relief funds as it chooses. The results reported in Table 3 assume the city allocates the new monies to additional public services. Under this assumption, Reform Policy 1 improves estimated city house values by $155 per house, but suburban house values are estimated to fall by $95. Neither effect should be considered economically significant. The total gain in wealth for regional home owners is a very modest $335 million, about 0.03 percent of their initial wealth, and all the gain goes to city residents.

Reform Policies 2 and 3. Reform Policies 2 and 3 look more promising. Under the second reform strategy, the city is required to allocate its poverty-relief funding to a 10 percent reduction in wage-tax rates for residents and nonresidents. The average city house is estimated to rise in value by $1087, or about 1.82 percent of its pre-reform value. But, again, the value of the average suburban house remains essentially unchanged, falling by $63. Overall, under Reform Policy 2, total home-owner wealth, regionwide, rises

[16] Chester County residents must contribute an additional 0.18 percent of county income, or $38.389 million; Delaware County an additional 0.43 percent of county income, or $59.313 million; Montgomery County an additional 0.22 percent of county income, or $50.988 million. The total additional contributions from the four suburban counties is $191.312 million. When these funds are given to Philadelphia for poverty relief, the city's net contribution to regional poverty spending becomes $342.880m - $191.312m = $151.568m, which is 0.60 percent of city resident income. For suburban counties as a whole, $191 million equals 0.29 percent of aggregate suburban income as reported in the 2000 census.

TABLE 3

Regional Financing for Regional Poverty

POLICY REFORMS	City Average House Value (% Change from Status Quo) (1)	Suburban Average House Value (% Change from Status Quo) (2)	Regional Total House Value (% Change from Status Quo) (3)
NO REFORM: STATUS QUO	$59,700 (-)	$157,836 (-)	$111.76 billion (-)
REFORM POLICY 1: "NO STRINGS ATTACHED"	$59,855 (0.26%)	$157,741 (-0.06%)	$111.99 billion (0.03%)
REFORM POLICY 2: UNIFORM WAGE TAX CUT	$60,787 (1.82%)	$157,773 (-0.04%)	$112.21 billion (0.50%)
REFORM POLICY 3: NONRESIDENT WAGE TAX CUT	$60,960 (2.11%)	$160,614 (1.76%)	$113.74. billion (1.87%)

Because all the required data for census year 2000 are not yet available, the estimates of the post-REFORM POLICY house values reported above were computed from the simulation model of the Philadelphia economy calibrated for the census year 1990 (see Technical Appendix). For the suburban counties as a whole, the required equalizing transfer of $191 million equals 0.29 percent of current suburban income; I have therefore scaled the required suburban contribution to 0.29 percent of the 1990 suburban incomes for all policy simulations. The percentage changes in city and suburban house values are computed using this scaled transfer. The estimated percentage changes in regional house values from the 1990 simulated economy (reported in parentheses above) are then multiplied by the actual 2000 census house values (reported here under NO REFORM: STATUS QUO) to give estimates of the new, post-REFORM POLICY house values.

an estimated $558 million, or about 0.5 percent, and again the benefits are concentrated in the central city.[17]

Under Reform Policy 3, however, all home owners in the region benefit, not just those in Philadelphia. Even though suburban residents send money to the city, suburban house values rise an estimated 1.76 percent. Why? The answer lies in the more efficient regional economy that follows from Policy 3's required reduction in the wage-tax rate for nonresidents. The economic effect of the city's nonresident wage tax is to increase city firms' labor costs roughly in proportion to the tax rate. By reducing that tax rate, Reform Policy 3 lowers labor costs in the city, which encourages city businesses to expand, more city jobs, and, most important for suburban residents, more low-cost city goods and services. In dollars, the estimated net gain to a typical suburban family in our simulated regional economy will be an improvement in house values of $2780, or about $140 a year assuming a 5 percent interest rate. Under Reform Policy 3, a Philadelphia area suburban family would "invest" $220 per year in higher county taxes but then benefit by saving $360 per year through their consumption of lower cost, city-produced goods and services. The net gain is $140 a year. Not a lot of money, perhaps, but a very nice rate of return!

Most important, regional financing of the cost of poverty is an opportunity for Philadelphia and the suburban counties to work together for the benefit of all residents of the Delaware Valley. If it is done correctly — for example, city poverty relief is exchanged for lower commuter tax rates for suburbanites — regional fiscal reform can be a true win-win, enhancing house values in the city and suburbs alike.

CONCLUSION: FISCAL COOPERATION FOR FINANCING POVERTY

There is much to recommend our region's decentralized system of

[17] The estimates in Table 3 for Reform Policies 1 and 2 are reassuringly similar to the estimates for suburban-to-city aid reported in Haughwout and Inman (2002) for their national sample. In their national sample, cities that share county functions with their suburban governments, such as Pittsburgh, have significantly higher average house values than do cities, such as Philadelphia, which pay for county functions on their own.

public finance. But when there are important economic interdependencies across local jurisdictions, fiscal cooperation, not fiscal competition, is required. One important economic interdependency, known as agglomeration economies, occurs within our central cities. This interdependency creates significant production efficiencies and allows valued product diversity, both of which benefit all residents of the economic region. Inefficient public finances in a city, however, can undo these economies as firms and households leave the city. On its own, Philadelphia has made significant progress toward efficient city budgeting since its 1990 fiscal crisis. Growth in city workers' compensation has been brought in line with annual rates of inflation, and the resulting savings in conjunction with productivity improvements have allowed balanced-budget reductions in the tax rates for wages and gross receipts.

The problem that remains is the city's disproportionate share of the region's responsibility for poverty spending, a burden it bears for historical and legal reasons. Regional financing of regional poverty will neutralize this threat to Philadelphia's productive efficiency, and the region as a whole will benefit. Reform can be implemented within the existing structure of state financing of county poverty spending; no new metropolitan government is necessary, nor is there any need for regionwide taxation. What will be required, however, is a commitment on the part of the city and the four suburban counties to work together. One promising option would provide city poverty relief in exchange for lower wage tax rates for nonresidents. Under this reform, city and suburban residents both gain, perhaps by as much as an additional $2.0 billion in regional house values, or $2250 per household.[18] The source of the gain is a more efficient Philadelphia economy, made possible by tax relief for suburban commuters. With city and suburban cooperation for regional poverty financing, we all win.

[18] The total difference in Reform Policy 3 minus the status quo (Table 3): $113.74 billion - $111.76 billion = $1.98 billion, or approximately $2 billion.

REFERENCES

Bartik, Timothy. *Who Benefits from State and Local Economic Development Policies?* Upjohn Institute, 1991.

Beardsell, Mark, and Vernon Henderson. "Spatial Evolution of the Computer Industry in the USA," *European Economic Review*, 43 (June), 1999, pp. 431-56.

Carlino, Gerald. "Productivity in Cities: Does City Size Matter?" Federal Reserve Bank of Philadelphia *Business Review*, November/December, 1987.

Ciccone, Antonio. "Agglomeration Effects in Europe," *European Economic Review*, 46, 2002, pp. 213-27.

Ciccone, Antonio, and Robert Hall. "Productivity and the Density of Economic Activity," *American Economic Review*, 86, 1996, pp. 54-70.

Freeman, Richard, and Robert Valletta. "The Effects of Public Sector Labor Laws on Labor Market Institutions and Outcomes," in Richard Freeman and Casey Ichniowski, eds., *When Public Sector Workers Unionize.* Chicago: University of Chicago Press, 1988.

Glaeser, Edward, and Joseph Gyourko. "Urban Decline and Durable Housing," NBER Working Paper 8598 (2001).

Glaeser, Edward, Hedi Kallal, José Scheinkman, and Andrei Shleifer. "Growth in Cities," *Journal of Political Economy*, 100, 1992, pp. 1126-52.

Haughwout, Andrew, and Robert P. Inman. "Fiscal Policies in Open Cities with Firms and Households," *Regional Science and Urban Economics*, 31, 2001, pp. 147-80.

Haughwout, Andrew, and Robert P. Inman. "Should Suburbs Help Their Central City?" *Brookings-Wharton Papers on Urban Affairs*, 2002, Brookings Institution, pp. 45-88.

Haughwout, Andrew, Robert P. Inman, Steven Craig, and Thomas Luce. "Local Revenue Hills: A General Equilibrium Specification with Evidence from Four U.S. Cities," NBER Working Paper 7603 (2000).

Henderson, Vernon, Ari Kuncoro, and Matt Turner. "Industrial Development in Cities," *Journal of Political Economy*, 103, 1995, pp. 1067-90.

Inman, Robert P. "How to Have a Fiscal Crisis: Lessons from Philadelphia," *American Economic Review*, 85, 1995, pp. 378-83.

Jacobs, Jane. *The Economy of Cities.* New York: Random House, 1969.

Marshall, Alfred. *Principles of Economics.* New York: Macmillan Company, 1890.

Rauch, James. "Productivity Gains from Geographic Concentration of Human Capital: Evidence from Cities," *Journal of Urban Economics*, 34, 1993, pp. 380-400.

Rosenthal, Stuart, and William Strange. "Geography, Industrial Organization, and Agglomeration," Syracuse University, Department of Economics, 2001.

Summers, Anita, and Lara Jakubowski. "The Fiscal Burden of Unreimbursed Poverty Expenditures," *Greater Philadelphia Regional Review*, 1997, pp. 10-12.

Voith, Richard. "Changing Capitalization of CBD-Oriented Transportation Systems: Evidence from Philadelphia: 1970-1988," *Journal of Urban Economics*, 33, 1993, pp. 361-76.

TECHNICAL APPENDIX

An Economic Model of Philadelphia and Its Suburbs

The predicted changes in Philadelphia and suburban house values reported in Table 1 Columns 3 and 4 and in Table 3 were computed using a general equilibrium model of household and firm location, investment, employment, and production within the Pennsylvania portion of the Philadelphia MSA (Bucks, Chester, Delaware, Montgomery, and Philadelphia counties). The model is able to predict city and suburban population, employment, firm production, government spending and taxation, worker wages, and finally home values, given an initial fiscal and demographic position for the city and its suburbs. The analysis also takes as given land available for the location of households and firms, both within the city and within the surrounding suburbs (Bucks, Chester, Delaware, and Montgomery counties).

The *household sector* consists of three broad classes of families — those whose adult head of household works and earns the region's competitive market wage, those whose head is a manager and earns an exogenous (outside the model) managerial wage set by the national market for managerial talent, and finally, those whose adult head is unemployed and classified as a family in poverty or over the age of 65. The Philadelphia region must offer all its workers and managers the same "standard of living" — what economists also call "utility" — as available elsewhere in the country. Working families can choose to live either in Philadelphia or in the suburbs. We assign managerial households to live in the suburbs, even though they may work in Philadelphia. Dependent households (those in poverty or over 65) are assigned to live in Philadelphia or the suburbs so as to match the actual MSA data. Dependent households do not move in response to fiscal or market incentives. If a household lives in Philadelphia, it receives the common level of city services and pays city taxes, either as property, wage, or sales taxation. If the family lives in the suburbs, it will receive the average level of suburban public services and it will pay the average suburban property tax rate. Managers who commute into Philadelphia are also assessed the city's nonresident wage tax, but the burden of this tax is shifted back onto the manager's city firm as an added cost of hiring a mobile managerial worker.

The *production sector* of the metropolitan economy consists of city firms that produce and sell a composite private good and suburban firms that retail the same private good but which must import that good either from city firms or from firms located outside the metropolitan area. The composite good should be viewed in the broadest sense to include all goods and services a family might purchase in the marketplace, from food to clothing to entertainment to legal services to health care. For suburban retailers, it is cheaper to import this composite private good from Philadelphia firms for two reasons. First, because of agglomeration economies, Philadelphia firms might be the low-cost producer. Second, suburban retailers are closest to Philadelphia producers so this saves on transportation costs. If the demand for the composite good by suburban residents exceeds the exports available from Philadelphia firms, the suburban retailers must import the private good from more expensive regional or national providers.

The *government sector* consists of a single central city, specified to approximate the finances of Philadelphia, and a single, all-encompassing suburban government specified to approximate the finances of an average local government plus school district in the four counties surrounding Philadelphia. Each local government produces one common public service and pays for the service using local taxes and intergovernmental transfers from state and federal governments net of payments for local debt and underfunded pensions. In addition, both the city and the suburb must meet required spending obligations for their low-income and age-dependent populations. City public services benefit city firms and city residents, while suburban public services benefit suburban retailers and suburban residents. For financing, Philadelphia uses a property tax, resident and nonresident wage taxes, and a gross receipts tax on sales by city firms within the city. The suburban government uses a local property tax. Both local governments are free to choose their own local tax rates and thus the level of local government spending on the public service.

The regional economy will be in *equilibrium* when all firms within the region earn the national competitive rate of return on invested capital, all households living in the region receive the national "standard of living" or "utility," managers earn their nationally competitive after-tax managerial salary, and the city and suburban governments choose locally balanced budgets. If firms in the Philadelphia region make more than the competitive rate of return, more firms move into the metropolitan area; if they make less, firms exit. When firms move into the region, the demand for land and labor increase, leading to a rise in land prices and worker wages. The increase in worker wages raises the standard of living for residents in the region, leading to regional population growth. Having more regional workers moderates the initial increase in wages but reinforces the initial rise in land prices. The opposite effect on wages and land prices occurs when firms leave the metropolitan area.

Firms will choose to enter the region when they are more productive and thus more profitable here than elsewhere. An important source of production efficiency will be city agglomeration economies. Efficient city firms are more profitable, attracting firms into the city. Land prices and city wages rise. Having more city firms also means more low-cost city output for export to the suburbs. Philadelphia suburbs are now more attractive. Suburban retailing output expands and this in turn increases the demand for suburban labor. Both effects raise suburban land values and suburban wages. In the end, improving central city agglomeration economies makes all residents living in the Philadelphia region richer.

A complete description of the simulation model can be found in the 2002 article by Haughwout and Inman.