Federal Reserve Bank of New York

December 2001 Volume 7 Number 3

Economic Policy Review

The *Economic Policy Review* is published by the Research and Market Analysis Group of the Federal Reserve Bank of New York. Articles undergo a comprphensive refereeing process prior to their acceptance in the *Review*. The views expressed in the articles are those of the individual authors and do not necessarily reflect the position of the Federal Reserve Bank of New York or the Federal Reserve System.

www.newyorkfed.org/rmaghome

Federal Reserve Bank of New York Economic Policy Review

December 2001
Volume 7 Number 3

Contents

Articles:

Public infrastructure investment may indirectly affect firm productivity and household welfare through its impact on the location of economic activity. Existing infrastructure policies encourage firms and households to move from dense urban environments to the surrounding suburbs. Nevertheless, several recent studies have suggested that the concentration of producers and consumers within cities results in "agglomeration economies" that are socially beneficial. In light of these findings, the author recommends the creation of infrastructure investment authorities that would have the power to select and finance projects that promote the overall well-being of a given region. Such authorities would most likely direct a larger share of infrastructure investment to the central cities.

Between 1995 and 1998, actual growth in compensation per hour (CPH) accelerated from approximately 2 percent to 5 percent. Yet as the labor market continued to tighten in 1999, CPH growth unexpectedly slowed. This article explores whether this aggregate "wage puzzle" can be explained by changes in the pay structure—specifically, by the increased use of employee stock options in the 1990s. The CPH measure captures these options on their exercise date, rather than on the date they are granted. By recalculating compensation per hour to reflect the options' value on the grant date, the authors find that the adjusted CPH measure accelerated in each year from 1995 to 1999.

Andrew F. Haughwout

Infrastructure and Social Welfare in Metropolitan America

- Infrastructure investment may indirectly affect firm productivity and household welfare through its impact on the *location* of economic activity.
- State infrastructure policies currently favor decentralization—the opening of new territory to development and the movement of firms and households from dense urban environments to the surrounding suburbs.
- Recent research, however, suggests that the clustering of producers and consumers in a given geographic area is economically and socially beneficial.
- In light of this research, institutional reforms that would change the management and direction of public infrastructure investment may be in order. Agencies authorized to choose and finance investments that promote regional well-being would most likely target more investment to central cities and less to the surrounding suburbs.

Public infrastructure is an important part of a well-functioning urban economy. Such infrastructure—defined here as publicly owned and maintained physical capital—has historically played a central role in allowing cities to grow by mitigating or reducing problems such as congested roadways, potholes, water-main breaks, and overcrowded schools. Yet while the benefit of some public works can hardly be disputed, a key policy issue is whether *additions* to our stock of public infrastructure provide overall benefits that exceed their costs.[1] That is to say, is the amount of infrastructure we have sufficient, or would we benefit from an increase? Another important question is, do our institutional structures promote efficient infrastructure investment decisions?

As these questions suggest, the status of urban public infrastructure is an important topic. Education and highway facilities are being stretched to their limits in fast-growing cities and suburbs, while concerns are being raised about the level and physical condition of public works in slower growing, older central cities.[2]

No doubt, public investment is an important function of government, and it is particularly crucial at the state and local level. In 1999, states and localities invested more than $210 billion in equipment, software, and structures (Table 1). By combining this amount with the nearly $43 billion in nondefense investments made by the federal government, we see that new gross public investment in 1999 exceeded a quarter-trillion dollars, or 2.7 percent of GDP. Moreover, the

Andrew F. Haughwout is a senior economist at the Federal Reserve Bank of New York.
<andrew.haughwout@ny.frb.org>

This work was funded by a grant from the Wharton Real Estate Center and results in part from research funded by the National Science Foundation while the author was at Princeton University. The author is grateful to Janice Madden, Erica Groshen, and two anonymous referees, whose comments have clarified the arguments. The views expressed are those of the author and do not necessarily reflect the position of the Federal Reserve Bank of New York or the Federal Reserve System.

stock of publicly owned nondefense capital in 1999 exceeded $4.5 trillion, or nearly 50 percent of GDP.[3]

Although complete data on the geographic distribution of this spending are not available, it is certain that a large share of these national totals, particularly the state and local portions, is going to public investment in and around America's metropolitan areas. More than 200 million people reside in these areas, and the public investments made there affect the lives of a large and growing share of the U.S. population.[4]

Accordingly, the question of whether we should increase the amount of infrastructure available has received much attention from economists. This article puts that research into a broad perspective, attempts to draw policy conclusions from what is known, and suggests some directions for further research.

Infrastructure investments can affect social welfare in two ways (see Appendix A). One way is by adding to economic growth. The relationship between infrastructure and economic growth has been the subject of intensive economic research over the past decade. The second way in which infrastructure investments can affect social welfare is by potentially improving the quality of life of those living in the invested area. For example, public parks, water systems, and other facilities can improve social welfare without having any effect on residents' incomes. This article also examines this second channel, which has received less attention in the research, in part because the value of quality-of-life improvements is difficult to measure.

Infrastructure investments may also indirectly affect social welfare in ways that have not been fully considered in public policy or research, such as location behavior, that is, where activities occur. For example, public works attract activities, such as the building of office parks near airports or housing developments near local roads. Furthermore, location behavior affects both economic growth and the quality of life.

Table 1
Nondefense Public Capital, 1999
Millions of dollars

	Gross Investment	Net Stock
Total	254,475	4,512,982
Federal	43,609	518,727
Equipment and software	32,738	105,496
Structures	10,871	413,231
State and local	210,866	3,994,255
Equipment and software	53,366	198,738
Structures	157,500	3,795,517
Buildings	71,380	1,582,955
Educational	38,762	789,027
Other buildings	32,618	793,928
Highways and streets	54,766	1,300,053
Conservation and development	2,360	69,710
Sewer systems	10,344	322,166
Water supply	7,796	212,826
Other structures	10,854	307,807
Memo:		
Total (percentage of GDP)	2.7	48.50
Federal	0.47	5.58
State and local	2.27	42.95
GDP	$9.3 trillion	

Source: U.S. Department of Commerce, Bureau of Economic Analysis.

Social Value of Public Investments: Direct Effects on Firms and Households

Investments in public infrastructure may influence society's general level of well-being in several different ways. In this section, we address what we refer to as infrastructure's direct effects.

Infrastructure's Value to Firms

The relationship between public investments and aggregate economic growth has been the subject of the bulk of recent infrastructure research by economists (Table 2 summarizes some of the findings). A storm of research in this area was touched off by a controversial study published by Aschauer (1989). Aschauer used *national* data for the postwar period to estimate the relationship between the nation's stock of public capital and aggregate income, or GNP. His results suggested that the marginal productivity of public capital—the addition to total income that could be expected from additions to the public capital stock—was enormous. Aschauer's most widely cited conclusion was that a 1 percent increase in the nation's infrastructure stock would raise aggregate output by 0.39 percent. This conclusion suggested that infrastructure was roughly twice as productive (at the margin) as private capital, and that the nation's public capital deficit was very serious. Aschauer went on to argue that a significant share of the national productivity growth slowdown that began in the early 1970s was because of declining rates of public investment.

Aschauer's study was immediately controversial. Because national time-series data tend to rise and fall together, skeptical researchers immediately focused on the statistical properties of his model (Hulten and Schwab 1991). Subsequently, the use of widely accepted statistical techniques eliminated the statistical significance of the time-series relationship between national infrastructure and economic growth (Aaron 1990). Criticism of the national approach quickly led to research that examined the effects of variation in public capital availability *within* the nation at the state or city level.

Like Aschauer, Munnell (1990) and Eberts (1986) estimated aggregate production functions, but used data for states and metropolitan areas, respectively. A 10 percent increase in private capital is usually assumed to raise output by about 1.5 to 2.0 percent. Munnell estimated that a 10 percent increase in public capital would raise private output by about 1.5 percent. This estimate was both economically and statistically significant and fueled further interest in the possibility of large unexploited returns to public investment. More recent refinements to the aggregate production approach have focused even more thoroughly on the model's statistical properties. In Holtz-Eakin (1994) and Garcia-Mila, McGuire, and Porter (1996), correction of the estimates for unobserved state-level characteristics reduces the output elasticity of public sector capital to zero. This suggests that the findings of Munnell resulted from correlations between infrastructure and unmeasured state traits. Another important study that used an analogous methodology found that although infrastructure had significantly positive effects on productivity, the price of new investment may ultimately exceed its benefits (Morrison and Schwartz 1996).[5]

Although there is some lingering controversy in the literature over whether infrastructure is, at the margin, productive at all in the aggregate, there seems to be general agreement that raising taxes to fund large additions to a particular state's stock of public works would not have very large positive effects on aggregate income in that state. This generally held view is important because it suggests that arguments for substantial increases in investment in state infrastructure must be based on something other than productivity improvements, such as infrastructure's value to household consumption.

Table 2
Selected Previous Estimates of the Value of Public Capital

Author	Data	Method	Estimated Output Elasticity	Use of Controls	Comments
Eberts (1986)	Standard metropolitan statistical areas over time	APF	0.04	No	Manufacturing only
Aschauer (1989)	National over time	APF	0.3 - 0.4	NA	Time-series data
Munnell (1990)	States over time	APF	0.15	No	
Holtz-Eakin (1994)	States over time	APF	~0	Yes	
Holtz-Eakin and Schwartz (1995)	States over time	APF	~0	Yes	Negligible interstate spillovers
Garcia-Mila, McGuire, and Porter (1996)	States over time	APF	~0	Yes	
Morrison and Schwartz (1996)	States over time	ACF	0.11	Yes	Aggregate cost function
Boarnet (1998)	California counties over time	APF		Yes	Negative intrastate spillovers
Rudd (2000)[a]	Standard metropolitan statistical areas over time	Compensating variations	0.03 - 0.1	NA	Cross-sectional data
Haughwout (forthcoming)[a]	Central cities over time	Compensating variations	0.0 - 0.03	Yes	Large household consumption effects

Note: APF is aggregate production function; ACF is aggregate cost function.

[a] The dependent variables in the compensating variations studies are local land and labor prices.

Infrastructure's Value to Households

The value of infrastructure as a contributor to quality of life has received little attention, in part because the value of unpriced goods is difficult to measure. However, the quality-of-life benefits of public investments are nonetheless likely to be very important because households are heavy users of infrastructure systems. Indeed, some public works are specifically designed to benefit households alone, and the justification for their construction does not point to their value in generating economic growth. One example is the construction of public parks and recreation facilities. This very significant line item in the nation's public capital budget seems clearly intended to provide direct benefits to households, with little regard for its effects on productivity.[6] Furthermore, the quality-of-life benefits of public works are not necessarily limited to this class of investments. Even elements of "core infrastructure," such as transportation and sewer and water systems, provide large direct benefits to households.

The problems inherent in measuring the quality-of-life benefits of public works have led to an overemphasis on relatively easy-to-observe productivity effects.

These quality-of-life benefits for households are excluded from productivity studies that currently dominate the literature. Consider the effects of a new road from your home to your place of work that cuts your one-way commuting time from one hour to thirty minutes. Will you arrive early at work each day, or sleep somewhat later? The treatment of public works in the productivity literature implies that employees will choose to arrive early at work, increasing their output. But at least some workers will likely claim at least a portion of the time for themselves, by eating breakfast with their families or reading the paper longer each morning. This potential for increased leisure will not be accurately measured in standard studies of income or productivity, but is a benefit nonetheless, as it improves the well-being of the individuals whose homes are affected by the new road. Accounting for the consumption value of public works is thus an important, but difficult, task.

Infrastructure investments may also *reduce* the availability of some quality-of-life benefits for household consumption. The classic example here is the often-argued case that investments in highway infrastructure induce more driving, which increases auto emissions and congestion, degrading the environment and consuming workers' valuable leisure time (see, for example, Downs [1994, p. 8]). Since air pollution and time lost to congestion are unpriced negative factors that are not traded in markets, any increase in them will not be directly reflected in incomes, but will reduce welfare.[7]

Few studies have tried to measure the consumption benefits of public investments on a large scale, but some evidence is available. Haughwout (forthcoming) uses a spatial equilibrium model to estimate the aggregate value that households put on public investments in central cities. In this model, households demand lower land prices and higher wages as compensation for living in locations with insufficient infrastructure. Using this method, Haughwout estimates that the present value to households of a $4.64 billion increase in central-city infrastructure is about $1.8 billion, far higher than the comparable benefit to firms. This estimate implies, however, that the aggregate benefit of such an investment is less than its cost, even when both household and firm benefits are included.[8]

The problems inherent in measuring the quality-of-life benefits of public works have led to an overemphasis on relatively easy-to-observe productivity effects. Whether these effects are large or not is only one part of the answer to the nation's public investment question. By themselves, findings that policies designed to raise taxes for public works do not significantly influence worker productivity cannot be taken as sufficient evidence that these policies have insignificant effects on social welfare. Until one presents convincing evidence of infrastructure's effects on household welfare derived from sources other than household income, the direct social welfare effects of infrastructure investments must be considered an open question.

Social Value of Public Investments: Induced Location Effects

There are many less direct ways in which infrastructure can affect economic growth and quality-of-life benefits. In this section, we focus on one of the most important ways and one that is central to policymaking in metropolitan areas: the relationship between infrastructure investments and the location of activities. The argument proceeds as follows: after reviewing evidence on the relationship between infrastructure and the geography of economic activity, we find that infrastructure does affect location behavior by changing the distribution of firms and households within metropolitan areas. Next, we outline the theoretical and practical conditions under which the location behavior of individual firms and households affects aggregate social welfare. In the presence of spatial externalities (such as traffic congestion or shared local public goods), the location choices of individual firms and households will have implications for society as a whole. We conclude that infrastructure investment that encourages decentralization may serve to undermine growth in productivity and social welfare.

Infrastructure and Intrametropolitan Location Patterns

The idea that public investments, especially in transportation systems, alter the intrametropolitan geography of economic activity is supported by both the majority of urban theory (Fujita 1989) and a substantial historical record (Jackson 1985; Tarr 1984). However, location has not been central to the most recent empirical work on infrastructure. Implicit in state infrastructure productivity studies is the notion that states with more public capital might grow faster than those with less. Yet few of these studies have taken seriously the possibility that marginal additions to infrastructure stock have their most important effects on *intrastate* patterns of activity.

> *Infrastructure is a productive and valuable good, but additions to its stock of public capital can actually reduce economic growth and social welfare, even before one considers the tax cost of new investments.*

Nonetheless, evidence from studies using a variety of methodologies indicates that although the *aggregate* (statewide) economic impact of new public works is small, infrastructure's local effects are much more significant. Here we attempt to draw new inferences from the numerous existing productivity studies and combine them with historical and contemporary evidence on the relationship between public works and the location of economic activity. Viewed from a geographic perspective, the seemingly conflicting results that emerge from studies that use different methodologies tell a consistent story. Moreover, the message that emerges is surprising: infrastructure is a productive and valuable good, but additions to its stock of public capital can actually reduce economic growth and social welfare, even before one considers the tax cost of new investments.

Most studies of the role of infrastructure in state-level economic growth agree that marginal increases in state public investment levels do not have large effects on aggregate output. The traditional interpretation of these results is that infrastructure's "productivity" is low or negligible. However, there is growing evidence for an alternative explanation—that infrastructure investment is a costly method of rearranging the economic geography of our metropolitan areas, with uncertain effects on productivity and welfare.

How Might Infrastructure Affect Location Behavior?

Public works will have important effects on location decisions if their benefits differ from one place to another. In the simplest framework, the benefits of a public good are the same irrespective of location. The usual example is national defense—a nuclear arsenal deters foreign attack everywhere in the nation. Most public works are fixed in place, however, and thus differ from this polar case—they provide the greatest benefits to those in a position to utilize them. This is the idea behind looking at the relationship between infrastructure and economic growth at the state level. However, the benefits from new public investments are unlikely to be uniform throughout a typical state because states are simply too large. Instead, infrastructure investments confer benefits on one part of a state relative to other parts, potentially influencing intrastate location patterns.

One of the principal reasons for building urban infrastructure is that it reduces some of the negative effects of urban life, allowing cities to grow and increase the productivity and consumption advantages they offer. In theory, public roads provide fast, cheap intrametropolitan transportation, public systems draw water to cities from places where it is abundant, and wastewater treatment plants and public landfills help dispose of city waste. In theoretical models, improvements to transportation and other local public goods increase land values in newly served areas, leading to higher density in those locations (Fujita 1989; Anas, Arnott, and Small 1998). A clear example from U.S. history was the development of urban trolley systems, which allowed workers to live farther from their workplaces, thereby expanding the size of the urbanized area (Jackson 1985; Margo and Atack 1998).

How Does Infrastructure Affect Modern Metropolitan Location Patterns?

Recent studies of the effects of public capital on intrametropolitan patterns of economic activity indicate that the place-specific effects of new public investments are substantial. These studies share the common assumptions that firms and households are mobile and that there is a free market for land. In these circumstances, the value of different locations will be reflected in local land and labor prices, and the marginal value of infrastructure investment may be calculated by comparing land values near where the investment took place before and after its completion.

Substantial academic and anecdotal evidence exists to support this intuition. Studies have found that land prices and infrastructure investments are positively related at the intrametropolitan scale (Voith 1993; McDonald and Osuji 1995; Haughwout 1997, 1999a). Evidence that intrastate patterns of activity are significantly influenced by infrastructure development is available at the county level for California (Boarnet 1998) and for the nation as a whole (Haughwout 1999b). Using a less formal approach, Garreau (1991) points out that the new agglomerations that he calls "edge cities" often arise near highway interchanges on the fringes of metropolitan areas. If the development of public capital stock has positive effects on some parts of state economies, one could question how it can have negligible effects on states in toto, as found in recent state-level productivity studies.

If location patterns affect productivity and household welfare, then infrastructure's effects on location patterns may be the most important way in which infrastructure influences well-being.

The answer is that the dominant effects of public investment must be on the location, not the level, of economic activity within states. For example, a given improvement in a state's highway system serves to move activity from its current location to newly accessible places elsewhere within that state. This conclusion is consistent with the notion that marginal increases in the nation's stock of public capital provide localities that receive new public works with an advantage compared with those places that get fewer public investment dollars. If correct, this description of infrastructure's benefits is very important because it implies that academic researchers have been looking in the wrong place in trying to understand infrastructure's effect on social welfare. It also means that, if location patterns affect productivity and household welfare, then infrastructure's effects on location patterns may be the most important way in which infrastructure influences well-being. As we explain below, recent evidence on the relationship between the location of activities and social welfare indicates that location patterns matter.

Location and Social Welfare

The idea that equilibrium location patterns in a market economy might have implications for social welfare depends on the existence of spatial externalities. If a household's or firm's choice of location has no effect on the well-being of others, then the land market will simply reflect the relative value of different locations and no relocation would improve a household's or firm's own welfare or society's welfare (Fujita 1989). However, the very existence of central business districts and cities implies that spatial externalities could be important and thus the location of activities may have effects on welfare.[9]

Businesses

A focus on the relationship between the location patterns of individual businesses and their productivity is an important theme in much of the recent literature on urban growth and development (Quigley [1998] and Anas, Arnott, and Small [1998] offer reviews of parts of this literature). Most of this research argues that the proximity of producers yields productivity benefits to these firms. The arguments in favor of these so-called agglomeration economies proceed on several fronts. First, geographically concentrated producers are believed to benefit from shared inputs. If, for example, an employee unexpectedly quits, a firm can find a replacement more easily if it is near other firms in a dense labor market. This proximity will reduce inefficient "down time" and will allow maximum use of the firm's private capital plant. Likewise, geographically clustered firms can share the cost and use of inputs such as those provided by producer service firms, which can operate at efficient scale when there are many potential consumers. Large concentrations of producers and consumers further allow for the sustainable production of a wider variety of goods and services than is available in smaller markets. A wider *variety* of available inputs allows producers to target their input purchases more precisely, promoting efficiency. A final major source of agglomeration benefits to producers is information spillovers. The idea that general and specialized information about products, processes, and markets circulates most freely in environments conducive to frequent formal and informal contact is not new (Marshall 1890; Jacobs 1969), but it has become increasingly important in recent discussions of the sources of economic growth (Lucas 1988; Ciccone and Hall 1996; Glaeser 1998).

Households

For households, an analogous set of arguments in favor of concentration applies. Households may benefit from the sharing of consumption goods that are best constructed on a large scale, such as public parks and stadiums. Households may value the insurance against unemployment that having multiple employers nearby offers, and they may benefit from the variety of private consumption goods that are sustainable in large agglomerations. Finally, household location may provide external effects analogous to the informational spillovers hypothesized to exist among firms.

Recent sociological and economic research indicates that segregation of poor minorities has significant social costs (Wilson 1987, 1996; Cutler and Glaeser 1997; Benabou 1996). When poor residents of urban areas are concentrated in particular neighborhoods removed from the middle class, those residents suffer from the lack of information about jobs, stable neighborhood associations, and access to high-quality public and private services (Wilson 1996). For poor children, segregation results in low-quality schools, relatively few working role models, and less absorption of mainstream social attitudes (Wilson 1987; Case and Katz 1991). Segregation thus makes it more difficult to escape poverty, contributing to higher social costs in both the short and long run.

Implications

These interactions among individuals and firms are reciprocal and not directly priced in markets. Individual firms both benefit from and contribute to the information that exists in dense agglomerations: although firms recognize the benefits they receive, their location decisions will ignore their benefits to others. Middle-class households' participation in local PTAs provides benefits to their own children, which they value, and to those of other parents, which they may not value. The reciprocal nature of these interactions means that individual location choices have effects outside the private benefit-cost calculus, and that the free market underprovides agglomeration. This opens up the possibility that atomistic location decisions will lead to socially inefficient patterns of business and residential locations.[10]

Infrastructure's Indirect Effect on Social Welfare

Even when public capital has value to both firms and households, its effect on social welfare will be ambiguous when location is also important. For example, county employment density is an important locational attribute of state economies: roughly speaking, states with jobs concentrated in a few counties appear to grow faster than those states in which employment is more dispersed (Ciccone and Hall 1996). The social welfare effect of infrastructure investment will partially depend on whether the investment encourages or discourages county employment density. The evidence reviewed above suggests that current state infrastructure policies serve to reduce density, which in turn implies that infrastructure's effect on location offsets at least part of its value to individual firms and households.

> *If infrastructure were ineffective in attracting firms to new locations, it would not be able to reduce agglomeration economies and slow productivity growth.*

The intuition is that new state public works serve primarily to open up new territory for urban development. Because public works are valuable and are disproportionately placed in relatively undeveloped areas (Voith 1998b; Haughwout 1999b), public investments provide individual firms and households with incentives to move from more dense to less dense environments. In making this decision, individual firms and households ignore their contribution to socially beneficial density and diversity. The result is too much decentralization from society's perspective. Thus, the irony of the effect of infrastructure investment is that it can reduce aggregate productivity and welfare, but only if it is productive for individual firms and valued by individual households. If infrastructure were ineffective in attracting firms to new locations, it would not be able to reduce agglomeration economies and slow productivity growth. A similar argument applies to households—infrastructure investment will induce them to relocate in socially damaging ways only if it is valuable to them individually.

Of course, decentralization, perhaps even excessive decentralization, can occur for reasons totally unrelated to public finance, including greater demands for space and income growth.[11] Infrastructure plays the role of accommodating these individual desires. If public works were free goods, there would be somewhat less reason for concern, since decentralization could arguably be the natural result of a market process.

> *Because individual firms and households consider only their private costs and benefits of decentralizing and ignore the negative effects on others, the private market is likely to produce too little density.*

Nonetheless, because individual firms and households consider only their private costs and benefits of decentralizing and ignore the negative effects on others, the private market is likely to produce too little density. Thus, social welfare might still be improved with careful bargaining over the level and location of infrastructure, even if public works were free.

However, public works are not free. Most infrastructure research using the aggregate production and cost function approaches has downplayed the cost of infrastructure investment and focused primarily on its benefits. If it can be established that improvements in the stock of public capital have little or no effect on the economy, regardless of their cost, then there is little reason to be overly concerned with how these investments are financed. We argue, however, that the distribution of the (financial and nonfinancial) costs of infrastructure investment is crucial to understanding the incentives for decentralization that they create.[12]

Selecting and Financing Urban Infrastructure Projects

The benefit principle of local public finance states that efficiency requires those who benefit from a particular public expenditure to pay for it. The finding that marginal increases in infrastructure stocks are associated with land-value benefits in places that receive new public works implies that efficiency may be realized if public investment decisions are made by a local institution armed with a land tax. Governments that tax land may apply a simple rule of thumb when evaluating proposed public works: raise taxes to make investments that will raise the price of land after accounting for both the cost of taxes and the benefits of the new investment (Brueckner 1979; Brueckner and Wingler 1984). As a first approximation, municipal governments seem to fit this description, and there is modest evidence that local government decisions are approximately efficient by this "local property value maximization" standard. But the existence of benefit and cost spillovers means that locally driven infrastructure policymaking is likely to lead to inefficiencies.

First, local governments in the United States tax *property*, not land. The taxation on property covers both capital and land. This raises a host of complex issues about who pays the property tax.[13] Because capital taxes reduce the national return on capital and are paid by owners of capital everywhere, part of the financial cost of local spending on public works is exported to owners of capital who live outside the jurisdiction that levies the tax. When part of the financial cost of investment can be exported through the tax system, local governments will rationally choose levels of spending in excess of the socially optimal amount.

The second major concern is that urban infrastructure investments appear to generate substantial benefits and costs that spill over municipal boundaries (Haughwout 1997, 1999a). If infrastructure investments induce relocations, then the decision of whether to make investments must be based on consideration of both the pluses and minuses. Ideally, the policymaking institution must be large enough to internalize all the relevant externalities. In urban areas, this would seem to suggest a regional authority, with a mandate to finance and make those investments that improve overall regional well-being. Such a body would not make investments that simply relocate activity from one place to another, unless the spatial reorganization somehow were to offer net benefits. Given the evidence on spatial externalities described above, relocations that increase density might meet this criterion.

However, there are no regional infrastructure investment authorities with the power to raise taxes to fund investment and allocate it according to regional benefit-cost principles. The closest approximations are metropolitan planning organizations (MPOs), which are empowered under federal law to plan and prioritize regional transportation investments. Among the criteria that MPOs are mandated to consider is metropolitan land use, which theoretically allows them to evaluate the effects of new investments on agglomeration economies and neighborhood externalities.

However, MPOs are imperfect instruments for planning infrastructure investments rationally. The institutional structure of these bodies typically makes them unrepresentative of the metropolitan area's population. (See Lewis [1998] and Appendix B for information on MPOs in the Second District.) In general, political jurisdictions within the planning area are equally represented, in spite of the fact that they have unequal numbers of residents. For example, the rule in the New York–northern New Jersey area is one county, one vote, although New York City, Jersey City, and Newark each have an additional vote. However, tremendous disparities remain: tiny Putnam County (population 95,000) has the same formal influence as Nassau County (population 1.3 million).

In addition, MPOs do not choose how much money to invest, only the distribution of funds over potential improvements (Boarnet and Haughwout 2000). Finally, the authority of these bodies is limited to transportation projects funded by state and federal governments, meaning that other kinds of public investment decisions, including locally funded road projects, may be made in a fragmented manner,

There are no regional infrastructure investment authorities with the power to raise taxes to fund investment and allocate it according to regional benefit-cost principles.

uncoordinated with transportation. Of course, it is valuable to match the geographic scale of the decision-making body with the area affected by its policies. The geographic area affected by transportation investments may be quite different from the area affected by sewer investments or public parks, suggesting that authorities with different service areas should make these different decisions. Yet the complementarities between transportation and other capital services suggest that the coordination of their planning would offer efficiency gains.

Finally, funding for local public works involves substantial intervention by higher levels of government, particularly at the state level. State decision makers might seem to have an appropriately broad perspective that allows them to evaluate benefits and costs that appear anywhere in the state, but in practice direct state infrastructure investments, especially for highways, appear to undermine rather than support dense urban environments (Haughwout 1999a, 1999b; Voith 1998). A better understanding of state infrastructure decision making is crucial to rationalizing the nation's urban infrastructure investment policies. For the time being, what is clear is that the prominent state role adds to the net local benefits of investment by reducing the cost.

Localities in many states depend primarily on property tax revenues to finance the local cost of public services. New jobs, and especially the commercial and industrial tax base that they bring with them, are attractive ways for localities to finance new services or reduce tax rates on resident homeowners (Danielson and Wolpert 1992). The benefits of new development thus give towns strong incentives to build new public works funded in large part by nonresidents. Those with the most to gain are the rapidly growing areas that tend to be most heavily represented on regional transportation planning boards, and they presumably use their disproportionate influence to bring in money for new roads. These localized benefits are financed by the rest of the population, through higher taxes and, perhaps, lower productivity growth as our cities are undermined.

Conclusion

Some responsibility for making infrastructure investment decisions rests with state governments as well as with local and regional authorities. Financing, meanwhile, is a complex combination of state and local taxes and user charges. A typical municipality, which will realize much of the benefit of a new investment but pay only a portion of the cost, has significant incentives to overinvest. Regional authorities, which could theoretically constrain the worst of these impulses, generally have limited power and overrepresent underdeveloped, sparsely populated areas. Given what we know about the effects of infrastructure and the importance of fostering dense urban environments, the reform of strategies for investing in urban infrastructure could significantly enhance social welfare. Accordingly, researchers might consider refining their estimates of the spatial scale at which agglomeration benefits operate. They might also relate these estimates more directly to different types of infrastructure investments. At the same time, research into the determinants of investment policies, particularly at the state level, seems warranted.

Our current state of knowledge supports a few tentative directions in which policy reform might go. The crucial elements in a reform package would be the creation of institutions capable of balancing the full costs and benefits of new investments and the allocation of direct state investments in such a way as to maximize their contribution to aggregate welfare. Often this will mean relatively more public investment in central cities and probably less in their surrounding suburbs. The difficulty lies in finding politically feasible changes to infrastructure policymaking strategies that will represent real improvements to the current system.

To date, attempts to eliminate state support for infrastructure investments outside of specifically designated areas have been controversial and have met with mixed success. Maryland's "Smart Growth" initiative has been highly touted, but it has been in place for only three years (Gurwitt 1999). In New Jersey, the state plan has led to conflict between state and local governments, as the latter continue to face increasing development pressures, giving them incentives to shift the financial and nonfinancial costs of infrastructure development onto nonresidents (Hamill 1992).

Local efforts to establish regional authorities with the power to make and finance a broad array of infrastructure investments have also met with mixed success. In a recent case, voters in Jefferson County, Alabama (which includes Birmingham and a portion of its suburbs), twice rejected proposals to increase sales taxes to fund a regional infrastructure agency that would have emphasized funding to the central city (Sweeney 1999). However, in Atlanta, where sprawl and pollution problems are among the worst in the nation, the state has created a powerful regional agency to manage and finance infrastructure investments (Ehrenhalt 1999). Calls for increased regional cooperation on infrastructure and land use have recently been heard in even the staunchly small-government-oriented areas of San Diego and Dallas (Murphy 2001; Michaels et al. 2001).

All told, the record is mixed, and in most metropolitan areas public investment decisions continue to be made on a case-by-case, locality-by-locality basis. Perhaps the most encouraging development is the apparent recognition among suburban voters that new infrastructure investments, particularly roads, do not seem to solve the problems associated with uncontrolled suburban growth (Egan 1998). Purchases of open space and regulation of new development are currently the most common responses to this problem. Nevertheless, it is a small step from this recognition to the conclusion that larger shares of infrastructure investment should be directed to maintaining and improving existing facilities, which are disproportionately found in densely developed central cities. Indeed, politicians, including former Vice President Gore (Egan 1998), and the popular press (*Philadelphia Inquirer* 1999) have begun to argue that the way to minimize the cost of growth is to direct growth into areas that have seen declines and that offer existing, albeit aging, infrastructure capacity.

Of course, it is likely no accident that the crucial suburban support for managing the location of growth and developing city infrastructure systems has occurred during a period of extraordinary prosperity. When traffic is among the issues that voters most often cite as important, it seems a safe bet that times are fairly good. In Seattle, growth management has seen support ebb and flow with the region's economic fortunes (Egan 1996). In Birmingham, whites living below the poverty line were the group most strongly opposed to the creation of a regional authority; wealthier suburbanites and African-Americans voted for it (Sweeney 1999). In San Diego and Dallas, the central city is making the case for regional institutions; it is unclear what the suburban reaction will be.

Just as the effects of today's investment decisions will be felt for decades to come, marginal adjustments to annual investment policies will provide little relief in the short run. Durable improvements in the nation's urban infrastructure policymaking require that new institutions with substantial staying power be created. Such institutions—to be maximally efficient—would accurately represent the region's population, have responsibility for planning and financing regional infrastructure investments, and work within an explicit mandate to evaluate the land-use changes that result from new investments. However, this ideal system will be difficult to achieve.

Perhaps the most important steps one could take in the short run are to encourage regional transportation planners to consider regionwide land-use effects and to support an open exchange of views among the affected parties. A good first step would be to create forums in which the regional implications of infrastructure investments could be discussed by all interested parties, equally represented.

Appendix A: A Mathematical Approach

The argument is shown mathematically by examining a social welfare function. This function provides an ordinal ranking of society's preferences over different outcomes:

$$W = \{WY(G), Z(G)\}\ . \tag{A1}$$

Society is made up of individuals; businesses do not play a role in the social welfare function except in that they provide incomes and otherwise affect the well-being of individuals. The idea in equation A1, then, is that social welfare (W) is determined in part by total income (Y), which allows consumption of private goods that are traded in markets, and in part by other factors that are not reflected in incomes (Z). The social value of increases in public investment (G) is determined by their effect on these components of social welfare. The effect of infrastructure on social welfare is given as:

$$\frac{dW}{dG} = \frac{\partial W}{\partial Y}\frac{dW}{dG} + \frac{\partial W}{\partial Z}\frac{dZ}{dG}\ . \tag{A2}$$

The components of infrastructure's value are represented by the terms on the right side of equation A2.

Infrastructure's contribution to income (which is also its net value as a factor of production) is given as $\frac{dY}{dG}$, and $\frac{\partial W}{\partial Y}$ translates income changes into changes in social welfare. If infrastructure has a positive effect on firms, then workers and owners of private capital will see income increase as public capital increases, meaning that $\frac{dY}{dG} > 0$. The social value of infrastructure as a contributor to the availability of unpriced consumption goods is $\frac{dZ}{dG}$.

If infrastructure has significant effects on the location of economic activity (L), then equation A1 must be rewritten as:

$$W = W\{Y(G, L(G)), Z(G, L(G))\}\ . \tag{A3}$$

Now, the effect of infrastructure growth on social welfare is given as:

$$\frac{dW}{dG} = \frac{\partial W}{\partial Y}\left(\frac{\partial Y}{\partial G} + \frac{\partial Y}{\partial L}\frac{dL}{dG}\right) + \frac{\partial W}{\partial Z}\left(\frac{\partial Z}{\partial G} + \frac{\partial Z}{\partial L}\frac{dL}{dG}\right) \tag{A4}$$

or

$$\frac{dW}{dG} = \frac{\partial W}{\partial Y}\frac{\partial Y}{\partial G} + \frac{\partial W}{\partial Z}\frac{\partial Z}{\partial G} + \frac{\partial W}{\partial Y}\frac{\partial Y}{\partial L} + \frac{\partial W}{\partial Z}\frac{\partial Z}{\partial L}\frac{dL}{dG}\ . \tag{A5}$$

On the right side of equation A5, the first two terms are positive if infrastructure is valuable to households and firms, and the signs of the last term will depend on the effect of infrastructure on location patterns. For simplicity, let us take L as a measure of the density. If density is a good from society's viewpoint (that is, if $\frac{\partial Y}{\partial L}$ and $\frac{\partial Z}{\partial L}$ are positive, as suggested in the literature), the social welfare effect of infrastructure investment will depend on whether it encourages or discourages density (that is, whether $\frac{dL}{dG} > 0$). The evidence reviewed suggests that current state infrastructure policies serve to reduce density ($\frac{dL}{dG} < 0$), suggesting that infrastructure's effect on location may at least partially offset its direct value to individual producers and households.

Appendix B: Metropolitan Planning Organizations in the Second District

In the Second Federal Reserve District, metropolitan planning organizations (MPOs) exist for all urbanized areas. Three groups in the New York City region share transportation planning authority. The New York Metropolitan Transportation Council (NYMTC) serves the city and the suburban counties of Rockland, Nassau, Suffolk, Putnam, and Westchester. The thirteen-county northern New Jersey region is served by the North Jersey Transportation Planning Authority (NJTPA). The South Western Regional Planning Agency (SWRPA) covers the eight municipalities of lower Fairfield County, Connecticut. The structure of the New York region's MPOs is fairly typical of those in other parts of the country.

NYMTC's voting members are the five suburban county executives, the New York City Planning Commission chairperson, the New York City Department of Transportation commissioner, the New York State Department of Transportation commissioner, and the Metropolitan Transportation Authority chairperson. The NJTPA Board of Trustees consists of one elected official from each of the fifteen subregions—the thirteen counties and two major cities, Newark and Jersey City. The Board also includes a governor's representative, the commissioner of the New Jersey Department of Transportation (NJDOT), the executive directors of New Jersey Transit and the Port Authority of New York and New Jersey, and a citizens' representative appointed by the governor. Only the Connecticut MPO's makeup is explicitly linked to population. Each of the eight Connecticut municipalities is represented by two members of SWRPA, with a town receiving an additional member for each 50,000 residents.

The structure of the NYMTC leads to radical differences in representation, with Putnam County's 95,000 residents having the same official influence as Nassau's 1.3 million. Although they have two representatives on the MPO, New York City residents are still underrepresented by this "votes per capita" measure. With 7.4 million residents, the city has one delegate per 3.7 million residents, a ratio far lower than any other ratio in the region. In New Jersey, the 857,000 residents of Bergen County have one vote on their MPO, as do the 125,000 residents of Hunterdon County. Because the mostly densely settled counties are generally those that are relatively close to New York City, for most of the area's residents their representation on the MPO depends on how far from New York City they live. In general, the rule is the farther away the county, the more representation its residents have in regional transportation planning. Suffolk County, New York, and Ocean County, New Jersey, both have relatively low per capita representation and are exceptions to this general rule.

The Structure of Representation of Two Second District MPOs

County or City	MPO	Distance from New York City (Miles)	MPO Votes per Million Residents
New York City	NYMTC	0.0	0.3
Hudson	NJTPA	6.1	1.8
Jersey City	NJTPA	6.5	4.3
Bergen	NJTPA	13.8	1.2
Essex	NJTPA	14.4	1.3
Newark	NJTPA	11.1	3.8
Union	NJTPA	19.3	2.0
Nassau	NYMTC	20.3	0.8
Passaic	NJTPA	24.7	2.1
Rockland	NYMTC	26.1	3.5
Westchester	NYMTC	28.6	1.1
Morris	NJTPA	30.7	2.2
Middlesex	NJTPA	32.6	1.4
Monmouth	NJTPA	35.0	1.6
Somerset	NJTPA	36.7	3.5
Sussex	NJTPA	45.3	6.9
Putnam	NYMTC	46.9	10.5
Hunterdon	NJTPA	51.5	8.0
Warren	NJTPA	53.7	10.0
Ocean	NJTPA	64.0	2.0
Suffolk	NYMTC	68.5	0.7

Notes: NYMTC is New York Metropolitan Transportation Council; NJTPA is North Jersey Transportation Planning Authority. Jersey City and Newark each have an additional representative on the NJTPA Board of Trustees.

Endnotes

1. Simply maintaining the current public stock costs tens of billions of dollars per year. Increasing the available infrastructure stock would require even more funding.

2. In 1999, for example, more than 50 percent of all cities considered infrastructure needs a top concern, while 68 percent reported increasing their infrastructure spending during fiscal year 2000. See Pagano and Shock (1999, 2000) for details.

3. Nearly 90 percent of the nation's nondefense public capital is owned by state and local governments.

4. In 1998, more than 80 percent of the U.S. population lived in metropolitan areas. See U.S. Census Bureau, Statistical Abstract 2000, Table 33, for details.

5. Readers interested in the infrastructure productivity literature should consult the complete reviews found in Gramlich (1994) and Eberts and McMillen (1999). For explicit and implicit critiques of the dominant aggregate approaches, see Haughwout (1998, forthcoming) and Rudd (2000).

6. State and local governments alone spent nearly $5.7 billion on parks and recreation capital in fiscal year 1997. We make the qualification here because secondary arguments in favor of park construction may mention that happy, healthy workers will be more productive, but we take these as comparatively unimportant justifications for these expenditures.

7. The fact that the value of time lost to congestion can be estimated does not mean that it is truly priced or that congestion reduces national income. Instead, these estimates reflect the value of leisure that drivers forgo because of congestion. Many land-use models predict that reductions in transportation costs will lead to increased residential decentralization (Anas, Arnott, and Small 1998), but whether these changes in residential patterns have in fact increased the length of commutes is more controversial. Gordon and Richardson (1994) argue that residential decentralization has not been accompanied by longer commutes, since employment locations have tended to decentralize as well.

8. The household present-value estimate of 39 cents per dollar of net investment (1.8/4.64 = 0.39) is not directly comparable to the Aschauer (1989) elasticity estimate of 0.39, which is an annual figure. Aschauer's result implies that public capital was dramatically undersupplied in the nation; the results in Haughwout (forthcoming) suggest an oversupply of public capital in large, older central cities.

9. A classic example of a positive spatial externality is the relationship between an apiary and an apple orchard. Both benefit from proximity to the other: the apple blossoms improve the quality of the honey the bees produce and the bees, in turn, help pollinate the apple trees. But if the beekeeper has an opportunity to move his hives to another location, he would consider only the net benefit to his honey business and would ignore any cost his move would have on the apple grower. Recent empirical evidence confirms the existence of very significant spatial externalities like these in both production and household welfare.

10. For individuals' unpriced consumption, relative proximity appears to provide the largest benefits and segregated neighborhoods seem to experience negative consequences. Unfortunately, less is known about the scale at which the relationship between density and welfare's productivity component operates. This question is important: without its answer, we cannot be certain whether a firm's move from the central city to an inner-ring suburb is a cause for concern. Identifying this scale is an important component of the research agenda outlined in our conclusion.

11. See Anas, Arnott, and Small (1998) for a comprehensive review.

12. Boarnet and Haughwout (2000) discuss the latter point with regard to highways; here the focus is on public investment more generally.

13. Hamilton (1975) argues that with optimal zoning, local property taxation is equivalent to a local benefit tax, but more recent work, summarized by Mieszkowski and Zodrow (1989), argues that the local property tax is partly a benefit tax and partly a tax on capital. In addition, some local governments use other taxes, whose statutory incidence varies (Haughwout et al. 2000).

References

Aaron, Henry. 1990. Discussion of "Why Is Infrastructure Important?" In Alicia Munnell, ed., Is There a Shortfall in Public Capital Investment? Federal Reserve Bank of Boston Conference Series no. 34: 51-63.

Anas, Alex, Richard Arnott, and Kenneth Small. 1998. "Urban Spatial Structure." Journal of Economic Literature 36, no. 3: 1426-64.

Aschauer, David. 1989. "Is Public Expenditure Productive?" Journal of Monetary Economics 23, no. 2 (March): 177-200.

Benabou, Roland. 1996. "Workings of a City: Location, Education, and Production." Quarterly Journal of Economics 108, no. 3: 619-52.

Boarnet, Marlon. 1998. "Spillovers and the Locational Effect of Public Infrastructure." Journal of Regional Science 38, no. 3: 381-400.

Boarnet, Marlon, and Andrew Haughwout. 2000. "Do Highways Matter? Evidence and Implications of Highways' Influence on Metropolitan Development." Unpublished paper, Brookings Institution.

Brooks, Nancy, and Anita Summers. 1997. "Does the Economic Health of America's Largest Cities Affect the Economic Health of Their Suburbs?" Wharton Real Estate Center Working Paper no. 263.

Brueckner, Jan K. 1979. "Property Values, Local Public Expenditure and Economic Efficiency." Journal of Public Economics 11, no. 2 (April): 223-45.

Brueckner, Jan K., and Thomas L. Wingler. 1984. "Public Intermediate Inputs, Property Values, and Allocative Efficiency." Economics Letters 14 (April): 245-50.

Business Week. 1997. "The Ties that Lead to Prosperity." December 15.

Case, Anne, and Lawrence Katz. 1991. "The Company You Keep: The Effect of Family and Neighborhood on Disadvantaged Youth." NBER Working Paper no. 3705.

Ciccone, Antonio, and Robert Hall. 1996. "Productivity and the Density of Economic Activity." American Economic Review 86, no. 1: 54-70.

Cisneros, Henry. 1996. "Regionalism: The New Geography of Opportunity." U.S. Department of Housing and Urban Development Cityscape (December): 35-54.

Cutler, David, and Edward Glaeser. 1997. "Are Ghettoes Good or Bad?" Quarterly Journal of Economics 112, no. 3: 827-73.

Danielson, Michael, and Julian Wolpert. 1992. "Rapid Metropolitan Growth and Community Disparities." Growth and Change 23, no. 4: 494-515.

Downs, Anthony. 1994. New Visions for Metropolitan America. Washington, D.C.: Brookings Institution.

Eberts, Randall. 1986. "Estimating the Contribution of Public Capital Stock to Metropolitan Manufacturing Production." Federal Reserve Bank of Cleveland Working Paper no. 8610.

Eberts, Randall, and Daniel McMillen. 1999. "Agglomeration Economies and Public Infrastructure." In Paul Cheshire and Edwin Mills, eds., Handbook of Regional and Urban Economics. Vol. 3. New York: North-Holland.

Egan, Thomas. 1996. "Urban Sprawl Strains Western States." New York Times, December 29.

———. 1998. "Dreams of Fields: The New Politics of Urban Sprawl." New York Times, November 14.

Ehrenhalt, Alan. 1999. "The Czar of Gridlock." Governing, May.

Fujita, Masahisa. 1989. Urban Economics. New York: Oxford.

Garcia-Mila, Teresa, Therese McGuire, and Robert Porter. 1996. "The Effect of Public Capital in State-Level Production Functions Reconsidered." Review of Economics and Statistics 78, no. 1: 177-80.

Garreau, Joel. 1991. Edge City: Life on the New Frontier. New York: Doubleday.

Gillette, Clayton. 2001. "Regionalization and Interlocal Bargains." New York University Law Review 76, no. 1 (April): 190-271.

Glaeser, Edward. 1998. "Are Cities Dying?" Journal of Economic Perspectives 12, no. 2: 139-60.

Gordon, Peter, and Harry Richardson. 1994. "Congestion Trends in Metropolitan Areas." In Curbing Gridlock. Washington, D.C.: National Academy Press.

Gramlich, Edward. 1994. "Infrastructure Investment: A Review Essay." Journal of Economic Literature 32, no. 3: 1176-96.

Gurwitt, Rob. 1999. "The State versus Sprawl." Governing, January.

Hamill, Sam M., Jr. 1992. "Beyond the NJ State Plan." <http://www.njfuture.org/HTMLSrc/beyond.html> (August 2001).

Hamilton, Bruce. 1975. "Zoning and Property Taxation in a System of Local Governments." Urban Studies 12, no. 2: 205-11.

Haughwout, Andrew. 1997. "Central City Infrastructure Investment and Suburban House Values." Regional Science and Urban Economics 27, no. 2: 199-216.

———. 1998a. "Aggregate Production Functions, Interregional Equilibrium, and the Measurement of Infrastructure Productivity." Journal of Urban Economics 44, no. 2: 216-27.

———. 1998b. "Local Capital Spending in U.S. Metropolitan Areas." Public Works Management and Policy 3, no. 1: 73-91.

———. 1999a. "Regional Fiscal Cooperation in Metropolitan Areas: An Exploration." Journal of Policy Analysis and Management 18, no. 4 (fall): 579-600.

———. 1999b. "State Infrastructure and the Geography of Employment." Growth and Change 30 (fall): 549-66.

———. Forthcoming. "Public Infrastructure Investments, Productivity and Welfare in Fixed Geographic Areas." Journal of Public Economics.

Haughwout, Andrew, Robert Inman, Steven Craig, and Thomas Luce. 2000. "Local Revenue Hills: A General Equilibrium Specification with Evidence for Four Cities." NBER Working Paper no. 7603.

Holtz-Eakin, Douglas. 1994. "Public Sector Capital and the Productivity Puzzle." Review of Economics and Statistics 76, no. 1 (February): 12-21.

Holtz-Eakin, Douglas, and Amy-Ellen Schwartz. 1995. "Spatial Productivity Spillovers from Public Infrastructure: Evidence from State Highways." International Tax and Public Finance 2, no. 3 (November): 459-68.

Hulten, Charles, and Robert Schwab. 1991. "Is There Too Little Public Capital?" Paper presented at the conference "Infrastructure Needs and Policy Options for the 1990s." Washington, D.C.: American Enterprise Institute.

Jackson, Kenneth. 1985. Crabgrass Frontier: The Suburbanization of the United States. New York: Oxford.

Jacobs, Jane. 1969. The Economy of Cities. New York: Vintage.

Lewis, Paul. 1998. "Regionalism and Representation: Measuring and Assessing Representation in Metropolitan Planning Organizations." Urban Affairs Review 33, no. 6: 839-53.

Lucas, Robert. 1988. "On the Mechanics of Economic Development." Journal of Monetary Economics 22, no. 1 (July): 3-42.

Margo, Robert, and Jeremy Atack. 1998. "Location, Location, Location! The Price Gradient for Vacant Urban Land: New York 1835-1900." Journal of Real Estate Finance and Economics 16, no. 2 (March): 151-72.

Marshall, Alfred. 1890. Principles of Economics. London: Macmillan.

McDonald, John, and Clifford Osuji. 1995. "The Effect of Anticipated Transportation Improvement on Residential Land Values." Regional Science and Urban Economics 25, no. 3: 261-78.

Michaels, Dave, Selwyn Crawford, and Louise Applebome. 2001. "Kirk Praises Call for Regional Plan: Leaders React to Report's Vision for City." Dallas Morning News, March 5.

Mieszkowski, Peter, and George Zodrow. 1989. "Taxation and the Tiebout Model." Journal of Economic Literature 27, no. 3: 1098-1146.

Morrison, Catherine, and Amy-Ellen Schwartz. 1996. "State Infrastructure and Productive Performance." American Economic Review 86, no. 5: 1095-1111.

References (Continued)

Munnell, Alicia. 1990. "How Does Infrastructure Affect Regional Economic Performance?" In Alicia Munnell, ed., Is There a Shortfall in Public Capital Investment? Federal Reserve Bank of Boston Conference Series no. 34: 69-103.

Murphy, Dick. 2001. "A Vision for San Diego in the Year 2020: A City Worthy of Our Affection." State of the City Address, January 8.

Pagano, Michael, and David Shock. 1999. "City Fiscal Conditions in 1999." Washington, D.C.: National League of Cities.

———. 2000. "City Fiscal Conditions in 2000." Washington, D.C.: National League of Cities.

Philadelphia Inquirer. 1999. "Gains and Losses: City and Suburb Are Feeling the Shift of People in the Region. Together They Can Stem the Tide." Editorial, July 2.

Putnam, Robert. 1993. "The Prosperous Community." The American Prospect 4, no. 13: 35-42.

Quigley, John. 1998. "Urban Diversity and Economic Growth." Journal of Economic Perspectives 12, no. 2: 127-38.

Rudd, Jeremy B. 2000. "Assessing the Productivity of Public Capital with a Locational Equilibrium Model." Board of Governors of the Federal Reserve System Finance and Economics Discussion Series no. 2000-23.

Shoup, Carl. 1989. "Rules for Distributing a Free Government Service among Areas of a City." National Tax Journal 42, no. 2: 103-21.

Sweeney, Donald. 1999. "Looking Back to the Future: An Analysis of the Metropolitan Agenda for Birmingham, Alabama." Unpublished paper, Princeton University.

Tarr, Joel A. 1984. "The Evolution of the Urban Infrastructure in the Nineteenth and Twentieth Centuries." In Royce Hanson, ed., Perspectives on Urban Infrastructure, 4-60. Washington, D.C.: National Academy Press.

Voith, Richard. 1993. "Changing Capitalization of CBD-Oriented Transportation Systems." Journal of Urban Economics 33, no. 3: 361-76.

———. 1998a. "Do Suburbs Need Cities?" Journal of Regional Science 38, no. 3: 445-64.

———. 1998b. "Transportation Investments in the Philadelphia Metropolitan Area: Who Benefits? Who Pays? And What Are the Consequences?" Federal Reserve Bank of Philadelphia Working Paper no. 98-7.

Wilson, William. 1987. The Truly Disadvantaged: The Inner City, the Underclass, and Public Policy. Chicago: University of Chicago Press.

———. 1996. When Work Disappears: The World of the New Urban Poor. New York: Knopf.

Hamid Mehran and Joseph Tracy

The Effect of Employee Stock Options on the Evolution of Compensation in the 1990s

- As the labor market tightened in 1999, the growth rate of compensation per hour (CPH) unexpectedly slowed.
- The decline in CPH may be attributed to the rapid increase in new employee stock option grants relative to the realization of options awarded before 1999.
- Employee stock options are captured in the CPH measure on the exercise date, not on the date granted, and the options' value can change considerably over the several years that can elapse between these dates.
- A recalculation of CPH that reflects the value of options on the grant date suggests no downturn in compensation growth in 1999.

From an economic standpoint, the 1990s were a remarkable period. On the one hand, the decade produced the longest-running U.S. expansion. On the other hand, a by-product of this continued economic growth was a sharp tightening of the U.S. labor market. This growing scarcity of available workers raised the concern that accelerating wage demands would develop, possibly leading to renewed inflation.

The 1990s were also noteworthy for the emergence of two "wage puzzles." The first puzzle is associated with the 1992-95 period, when nominal compensation per hour (CPH) growth declined *and* the unemployment rate fell rapidly (Chart 1).[1] One explanation for this occurrence is that "worker insecurity" early in the expansion accounted for the tepid pay demands during this period.[2] From 1995 to 1998, the puzzle ceased to exist, as compensation growth accelerated and the unemployment rate fell below the 4 percent barrier. However, the second wage puzzle appeared in 1999, when compensation growth fell back below the 5 percent level despite continued labor market tightness during the year.

In this article, we examine the wage-puzzle phenomenon of the 1990s. Specifically, we explore whether changes in pay structure can account for the behavior of CPH during the decade. Labor markets have changed considerably over the past twenty years: workers today receive a higher portion of total

Hamid Mehran is an assistant vice president and Joseph Tracy a vice president at the Federal Reserve Bank of New York.
<hamid.mehran@ny.frb.org>
<joseph.tracy@ny.frb.org>

The authors benefited from discussions with Elizabeth Keating and Alan Marcus and from the comments of seminar participants at the Department of Labor and at Georgetown University Law Center's conference "Contracts with Highly Skilled Workers." They thank John Campbell, John Graham, and Martin Lettau for providing data, Nathanial Baum-Snow and Dan Burdick for excellent research assistance, and two anonymous referees. They also thank Hewitt Associates for technical assistance and summaries of its data and Thermo Electron for information on its employee stock options program and reporting. The views expressed are those of the authors and do not necessarily reflect the position of the Federal Reserve Bank of New York or the Federal Reserve System.

compensation in such nontraditional forms as profit sharing and stock options.[3] CPH captures profit-sharing payments and stock option realizations. However, employee stock options are reflected in total compensation on the date they are exercised—not on the date granted—and several years can elapse between these dates. Accordingly, the growing use of these stock options could be affecting the time that tight labor markets are reflected in CPH growth.

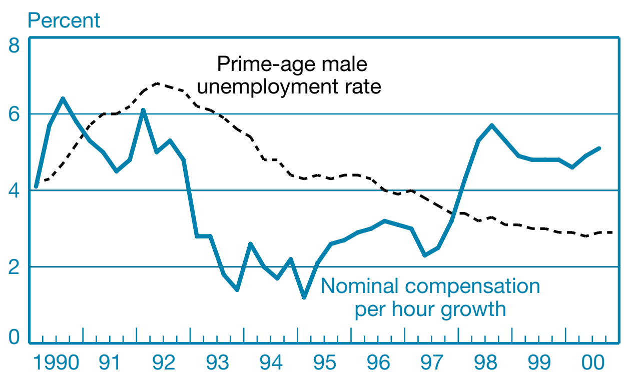
By analyzing the existing data, we determine how CPH growth is affected by the use of employee stock options. However, given the limitations of these data, we focus primarily on the second wage puzzle, that of the late 1990s. We find evidence that employee stock options may have had an appreciable impact on CPH during this period. In particular, when we recalculate compensation to reflect current stock option grants—rather than current realizations—we conclude that there was likely no downturn in CPH growth in 1999.[4]

The article is organized as follows. We begin by describing the essential institutional details of employee stock options necessary for our empirical work. We then discuss empirical models of stock option grant and realization decisions. Next, we use these estimates to assess the effect of stock options on compensation per hour. We conclude by addressing some general labor market implications of stock options.

Employee Stock Options

Employee stock options are the right to purchase a given number of shares of company stock at the "strike" price

Chart 1
Labor Market Tightness and Nominal Compensation per Hour Growth



Source: U.S. Department of Labor, Bureau of Labor Statistics.

between the vesting date and the expiration date of the options.[5] The vesting period is the interval between when a company grants the option and when the employee can first exercise the option. If the current market price for a vested option exceeds the strike price, the option is "in-the-money." If in-the-money options are exercised (that is, if the employee decides to purchase the underlying shares), the gain to the employee is the difference between the current market price and the strike price multiplied by the number of shares exercised.[6] If the current market price for a vested option is below the strike price, the option is "out-of-the-money." Although out-of-the-money options have no current value if exercised, they still have positive "option value," which reflects the possibility that the future market price of the stock may rise above the strike price prior to the options' expiration date.

Employee stock options can be structured either as incentive stock options or as nonqualified stock options. Incentive stock options must satisfy certain restrictions defined by the Internal Revenue Service that do not apply to nonqualified options.[7] The primary advantages to employees exercising incentive stock options are that the income derived is taxed as a capital gain, rather than as ordinary income, and the tax is levied when the underlying shares are sold, rather than when the option shares are exercised. Offsetting these tax gains to employees, however, is the loss of a tax deduction to the firm. In contrast to incentive stock options, the income gain from nonqualified stock options is treated for tax purposes as ordinary income to the employee as of the exercise date, and the company can deduct this cost as a labor expense. Employers are required to file quarterly reports (ES202s) that list all taxable sources of income paid to their workers, including realized nonqualified stock options. The ES202 reports are used as an input into total compensation. However, these reports do not break out the gain from nonqualified stock options from other sources of compensation. Nonqualified options became the dominant type of employee stock option following the reduction in marginal income tax rates in 1986.

In January 1993, the Securities and Exchange Commission (SEC) began requiring public firms to disclose in their proxy statements both the level of stock option grants to, and the option exercise activity of, their top five executives. The SEC also required companies to report their executive compensation for the two previous years in their annual filings with the agency. Beginning in 1991, then, it is possible to collect detailed information on public company stock option programs for top management. Although firms can value these option grants using any pricing methodology, the dominant method used is the Black-Scholes pricing formula.

Employee stock options differ in many important ways from traded stock options. Most notably, they are nontradeable. An employee can exercise a vested in-the-money option but cannot sell the option to an investor. An implication of this is that both the employee's valuation of the option and the timing of the exercise decision are affected by the employee's risk tolerance. An employee with a significant amount of wealth tied up in company stock options has a strong interest in diversifying the risk from movements in the value of the company stock. With traded stock options, the employee could simply sell some options in the market to another investor, an action that transfers but does not diminish the options' underlying value. With employee stock options, the employee would have to exercise the options in order to diversify his risk.[8] This creates an incentive for the early exercise of the options, which reduces their overall value because the employee forgoes the remaining option value. Huddart and Lang (1996) show that workers tend to exercise employee stock options soon after their vesting dates, and that this early exercise sacrifices roughly half of the value implied by the Black-Scholes pricing methodology (which is designed to price a traded stock option).

> *An employee with a significant amount of wealth tied up in company stock options has a strong interest in diversifying the risk from movements in the value of the company stock. . . . This creates an incentive for the early exercise of the options, which reduces their overall value.*

Employee stock options differ from traded stock options in two other key ways. As we observed, employee stock options are subject to vesting requirements and tend to have a significant time period until expiration. A variety of vesting schedules are used in practice, with the majority of plans incorporating vesting over two to five years.[9] In addition, an employee must exercise any vested in-the-money options prior to leaving the firm; any nonvested or out-of-the-money options must be forfeited upon termination of employment. This restriction creates an additional reason for early exercise of these options.[10]

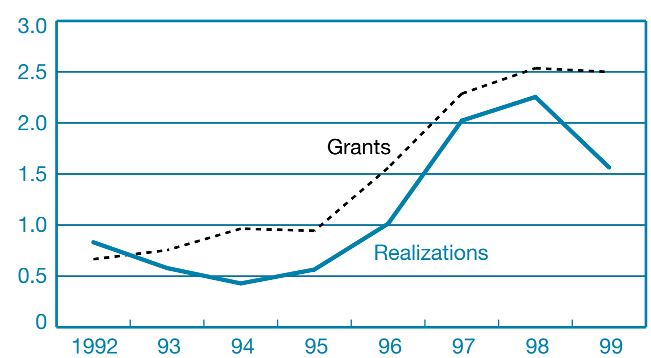
Measuring the Importance of Employee Stock Options

Our primary data source is Standard and Poor's ExecuComp database. ExecuComp includes annual data from proxy statements for the five highest paid executives in three cohorts of firms: the S&P 500, the S&P MidCap 400, and the S&P Small Cap 600.[11] Standard and Poor's makes some adjustments to the firms in the database each year. Our pooled sample, which covers the 1992-99 period, comprises a total of approximately 2,000 companies. We make extensive use of three specific items from the ExecuComp data: the total number of new grants to all employees, the number of grants and their value going to the top five executives, and the value of options exercised by the top five executives. We calculate the total value of all new grants in a year by scaling up the value of the grants to the top five executives by the ratio of the total number of options granted to the number of options granted to the top five executives.[12]

The ExecuComp data are valuable for examining general trends in the use of employee stock options during the 1990s. For example, over the decade, stock options became the dominant component of an executive's compensation package. We illustrate this remarkable change using two measures of the relative importance of executive stock options. The first is the ratio of the grant value of new options in a year divided by the executive's base salary and cash bonus. The second is the ratio of the income gain from stock option realizations in a year divided by the executive's base salary and cash bonus. Chart 2 presents the averages for these two ratios from 1992 to 1999.

Chart 2

Ratios of Stock Option Grants and Realizations to Top Five Executives' Base Salary and Cash Bonus



Source: Standard and Poor's ExecuComp database.

Early in the 1990s, both ratios indicated that stock options typically were smaller than an executive's base salary and cash bonus. By 1996, both ratios equaled or exceeded 1. Two years later, continued rapid growth in the expansion of executive stock option programs had pushed both ratios above 2, with new grants averaging around 250 percent of an executive's base salary and bonus. In 1999, the grant ratio leveled off, while there was a sharp reduction in the realization ratio.[13]

An important related question is whether the use of stock options is also filtering its way down the ranks of company pay structures. The ExecuComp data allow us to track the percentage distribution of total new stock option grants awarded to the top five executives. Although this is a very restrictive view of the diffusion of stock options down the corporate ranks, it has the advantage of providing some sense of recent trends. Chart 3 shows the equally weighted twenty-fifth, fiftieth, and seventy-fifth percentiles of these top five percentages from 1992 to 1999. Despite the dramatic rise in the use of stock options for executives, there has actually been a slight decline in the fraction of new stock option grants directed toward upper management. This indicates that there has also been a commensurate increase over the 1990s in the use of stock options for employees below the top management level.

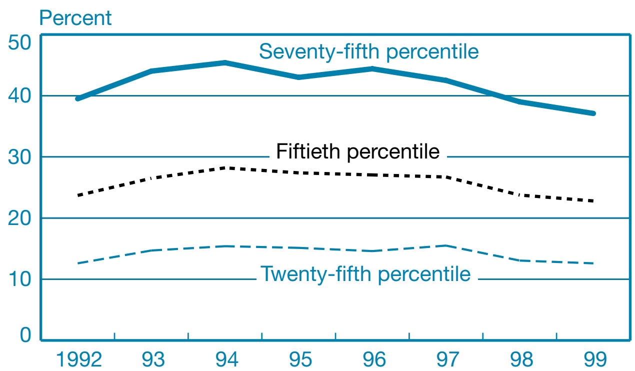
Given the rapid rise in the use of stock options, it is interesting to speculate on the effect of these options on aggregate compensation growth in the private sector. As noted earlier, aggregate compensation reflects nonqualified stock options when they are realized, rather than when they are granted. Unfortunately, there currently are no collected data that permit the direct measurement of the total size of stock option grants or realizations in the labor market. The alternative is to estimate total stock option grants and nonqualified stock option realizations by year. The growth rate of CPH net of the income from stock option realizations can then be constructed and contrasted with its actual growth rate. In addition, the cash value of new stock option grants can be added into this net-of-realizations CPH measure to arrive at a more accurate measure of current labor market pay conditions. We now turn our attention to implementing this approach.

The private, nonfarm sector consists of publicly traded and private firms. Over the past five years, public firms have accounted for between 47 and 50 percent of employment in the private, nonfarm business sector. The ExecuComp data consist entirely of publicly traded firms, and in 1998 the data covered roughly 46 percent of total employment in public firms.[14] Detailed characteristics of publicly traded firms are available from the COMPUSTAT data, and equity returns for these firms are available from the Center for Research in Security Prices (CRSP) data. For private firms, we have no similar data on their characteristics, nor do we have any details of stock option plans from which to draw any inferences. However, a recent Bureau of Labor Statistics study found that the incidence of stock options in privately held firms in 1999 was significantly below that for publicly held firms.[15] Based on this evidence, we focus our analysis exclusively on public firms.

There has actually been a slight decline in the fraction of new stock option grants directed toward upper management. This indicates that there has also been a commensurate increase over the 1990s in the use of stock options for employees below the top management level.

The basic question, then, is how best to use the ExecuComp data to estimate total stock option grants and realizations for publicly held firms. The simplest approach would be to assume for each year that all employees in these firms that are outside the ExecuComp sample are awarded new stock option grants and realize vested stock options at the average rate observed in that year for employees covered in the ExecuComp data. This approach, however, ignores potentially important variations across firms in their use of stock options that relate to firm

Chart 3
Percentage of Stock Option Grants to Top Five Executives



Source: Standard and Poor's ExecuComp database.

characteristics. Taking these variations into account may provide a more accurate estimate of the overall impact of stock options on total compensation.

Determinants of Stock Option Grants

We begin with the problem of estimating stock option grants, since the volume of prior grants is likely to be an important predictor of current realizations. For firms in the ExecuComp sample, we can calculate the Black-Scholes value of the total employee stock option grants made in the year. Although we are interested in understanding the determinants of a firm's decision regarding the total amount of new grants to make in a year, we find it useful to look at the executive compensation literature for guidance on an appropriate empirical specification.

The literature on executive compensation starts with the premise that optimal compensation policies should address agency problems between the firm's managers and its equity and debt holders. The two methods for ameliorating these agency problems are monitoring and incentives.[16] A general prediction is that stock options will be more extensively used when agency costs are high and monitoring is difficult. In addition, the accounting treatment of stock options discussed earlier suggests that firms may also use the options for tax or liquidity reasons.

We include several variables to control for expected agency costs. Monitoring may become difficult when a firm has significant growth opportunities. Information asymmetries may arise from these opportunities, making evaluation of the managers' investment choices more difficult (see, for example, Mehran [1992]; Smith and Watts [1992]; and Bizjak et al. [1993]). Stronger incentives therefore are needed to compensate for the monitoring difficulties. These additional incentives can be provided by increasing the share of stock options in total compensation. We measure a firm's growth opportunities using its market-to-book value. The prediction is that stock option grants will be positively related to this value.

It is also difficult to monitor managers in an environment in which a significant amount of noise is associated with the firm's performance (Lambert and Larcker 1987). In such an environment, a higher pay-performance sensitivity is warranted. Yermack (1995) proxies this sensitivity using the ratio of the relative variability of accounting returns versus stock returns. We focus just on the variability of stock returns over the prior year. The prediction is that higher stock return variability will lead to increased use of stock options. However, higher stock return variability also increases the manager's risk exposure, which should lead to a higher risk premium to compensate the manager for this added risk.[17] This risk premium increases the relative price to the firm of using stock options versus cash compensation, which may induce the firm to substitute away from stock options in its pay structure. The overall effect of stock return variability on the use of stock options therefore is ambiguous.

Capital structure may also exert an important influence on a firm's compensation system. Stock options, by increasing managers' pay-performance sensitivity, may encourage managers to pursue riskier investment strategies that tend to favor equity holders over debt holders. If this shift in investment strategies is anticipated by bondholders, the increased reliance on stock options will give rise to a debt premium that differentially impacts highly leveraged firms (John and John 1993). To reduce this agency cost of debt, highly leveraged firms may choose to scale back their use of stock options. This should lead to an inverse relationship between a firm's leverage and its reliance on stock options.[18]

To help control for any firm life-cycle effects on the use of stock options, we control for a firm's age, which we measure as the number of years over which the firm's stock has been traded. If young firms tend to be more cash-flow constrained, then we would expect them to rely more heavily on stock options. When a firm issues new stock options, it typically incurs no current expense, rather, the expense is shifted to the future, when the stock options are realized. Workers, however, value these new stock options, and are willing to accept lower current cash compensation as a consequence. This should lead to a negative relationship between firm age and the granting of stock options. We also directly proxy for cash-flow constraints using an indicator variable for whether the firm has a net operating loss in the current year.

The use of stock options varies with firm performance and firm size. Firms with a high return on assets tend to grant fewer new stock options. . . . Stock option grants tend to increase with firm size as measured by employment and total assets.

Our remaining firm-specific variables include measures of recent performance and size. We measure firm performance using return on assets. We use the size of the firm's assets and employment to control for possible scale effects. Finally, we include two-digit industry effects and year effects to control for any remaining differences across industries and time in the pattern of stock option grants.

Table 1 presents our estimation results for stock option grants (summary statistics are provided in Appendix A). For most of our control variables of interest, we divide the range of the variable into quartiles and create indicator variables for the upper three quartiles. The coefficient on an indicator variable should be interpreted as the difference in the use of stock option grants between a firm with a value of the indicated variable in the specified quartile and a similar firm with a value of the indicated variable in the bottom quartile (holding all values for other variables at their sample means).

The use of stock options varies with firm performance and firm size. Firms with a high return on assets tend to grant fewer new stock options. For example, grants for firms performing at or above the median in return on assets tend to be around 35 to 40 percent below the poorest performing firms. Stock option grants tend to increase with firm size as measured by employment and total assets. The employment relationship applies only to the top size quartile, while the asset relationship holds throughout the size range and is quite large in magnitude, but is imprecisely estimated. Core and Guay (1999) find that executive stock and option incentives are positively related to firm size as measured by equity value.

Empirical support exists for the agency cost of debt constraint on employee stock options. Controlling for other factors, we find that highly leveraged firms tend to pay out fewer new stock option grants. Firms in the highest quartile of leverage have new grants that are on average 26 percent below the level of firms in the lowest quartile of leverage. These results are in contrast to Yermack's (1995) empirical findings.

Monitoring problems arising from potential market opportunities and noisy environments also play an important role in determining the flow of new stock option grants. Firms with higher market-to-book value tend to have much more aggressive stock option programs, as evidenced by consistently higher flows of new stock option grants. This effect is especially pronounced for firms in the top market-to-book-value quartile that are predicted to have on average a 300 percent larger flow of stock option grants than firms in the bottom quartile. Higher stock return volatility reduces the magnitude of a firm's stock option grants. These findings are consistent with the existing empirical literature (see Core and Guay [1999, 2000]).

We also find support for the prediction that firms facing cash-flow constraints substitute stock options for cash compensation. The data suggest that firms experiencing a net operating loss in a given year have stock option grants that are 24 percent higher than those of similar firms with operating profits.[19] In addition, younger firms tend to rely more heavily on the use of stock options in their compensation structure: a ten-year increase in firm age is associated with an 8 percent decline in stock option grants.

Table 1

Determinants of Stock Option Grants

Variable	Percentage Change	Variable	Percentage Change
Return on assets		Market-to-book value	
Second quartile	-27.7**	Second quartile	43.4**
	(5.1)		(6.0)
Third quartile	-34.4**	Third quartile	104.8**
	(4.7)		(9.3)
Fourth quartile	-39.1**	Fourth quartile	300.3**
	(4.5)		(20.5)
Log employment		Stock return risk	
Second quartile	5.3	Second quartile	-13.5
	(15.6)		(18.0)
Third quartile	3.3	Third quartile	-31.7**
	(15.5)		(10.2)
Fourth quartile	64.7**	Fourth quartile	-23.3**
	(25.2)		(11.2)
Log assets		Net operating loss	24.2**
			(5.0)
Second quartile	89.1	Firm age (ten years)	-8.4**
	(94.4)		(1.5)
Third quartile	343.1	Number of observations	8,182
	(221.8)		
Fourth quartile	1,453.7*	R^2	0.46
	(779.7)		
Leverage			
Second quartile	-5.7		
	(4.5)		
Third quartile	-16.7**		
	(4.0)		
Fourth quartile	-26.6**		
	(3.8)		

Source: Authors' calculations, based on data from Standard and Poor's ExecuComp database and COMPUSTAT.

Notes: Ordinary least squares estimates of the percentage changes in the Black-Scholes grant value are reported with standard errors in parentheses. Two-digit industry and year effects are included in the specification.

** Statistically significant at the 5 percent level.

* Statistically significant at the 10 percent level.

Determinants of Stock Option Realizations

The ExecuComp data report the stock option realizations for each of the top five executives. What is not reported are the total stock option realizations generated by the other employees. To estimate total realizations, we assume that the time pattern of stock option realizations by the top five executives can be used to proxy for the time pattern of realizations by the remaining employees.[20] Specifically, we calculate a firm's total realizations in the year by scaling up the realizations by the top five executives using the ratio of total grants to top five grants from two years earlier.[21]

The empirical specification for a firm's total stock option realizations is motivated in part by the characteristics of employee stock option plans. In any given year, an employee has the right to realize any vested stock options. As previously noted, there is a strong tendency for employees to exercise options close to their vesting dates. While vesting schedules vary across firms, the typical vesting rules imply that it would be important to control for stock option grants from two to five years prior to the current year. Given the short time span covered by the ExecuComp data, we compromise and include only grants from two years prior to the current year.

Vested stock options are exercised only if they are in-the-money. Since the option strike price typically is set equal to the market price on the grant date, the cumulative stock return during the vesting period will determine whether an option is in-the-money on the date it vests. If a firm makes grants to employees over several years and uses a staggered vesting schedule, the appropriate stock return to examine would be a weighted average of different cumulative returns over the various vesting periods. Since we lack the detailed data necessary to calculate this particular stock return, we use as our proxy the firm's cumulative stock return over the prior two years.

A prominent feature of stock option realization data is that in a given year many firms experience no realizations, even if these firms have continuously made grants over the past several years. In our sample of approximately 5,189 firm/year observations for which we have complete data for all of our control variables, 32 percent involve no realizations by the firm in that year. To account for the high frequency of zero realizations in the data, we use a generalized Tobit specification. (Details on the Tobit model are provided in Appendix B.)

The generalized Tobit results are presented in Table 2. For ease of interpretation, we convert the generalized Tobit coefficients into three marginal effects: the implied impact of a variable on 1) the probability that a firm will experience a positive realization in the year, 2) the percentage change in the expected log realizations conditional on a positive realization, and 3) the percentage change in the unconditional expected log realizations.

Table 2
Determinants of Stock Option Realizations

Variable	Probability of a Positive Realization	Expected Realizations Conditional on a Positive Realization (Percent)	Expected Unconditional Realization (Percent)
Grants, lag two years	2.1**	75.1**	75.3**
	(0.6)	(5.3)	(8.5)
Cumulative two-year stock return			
Second quartile	18.0**	4.0	280.7**
	(2.5)	(4.1)	(89.9)
Third quartile	28.2**	89.9**	1,153.0**
	(2.5)	(22.9)	(328.6)
Fourth quartile	32.9**	205.3**	2,541.0**
	(2.5)	(37.3)	(750.3)
Log employment			
Second quartile	20.3**	-38.5	274.4
	(9.1)	(27.7)	(310.1)
Third quartile	23.2**	-25.5	432.7
	(8.9)	(32.9)	(432.5)
Fourth quartile	27.5**	-14.9	715.4
	(8.9)	(37.6)	(666.2)
Leverage			
Second quartile	-0.3	7.4	2.9
	(2.5)	(11.3)	(23.5)
Third quartile	2.3	12.9	31.1
	(2.5)	(11.8)	(29.1)
Fourth quartile	-2.0	2.2	12.7
	(2.7)	(11.9)	(21.4)
Market-to-book value			
Second quartile	11.9**	59.0**	221.5**
	(2.1)	(14.5)	(63.4)
Third quartile	21.7**	118.1**	772.8**
	(2.2)	(21.4)	(204.6)
Fourth quartile	26.5**	223.1**	1,653.2**
	(2.3)	(34.8)	(467.8)
Stock return risk			
Third quartile	5.9**	11.4	77.4**
	(2.1)	(10.4)	(39.3)
Number of observations	5,189[a]	3,508	5,189[a]

Source: Authors' calculations, based on data from Standard and Poor's ExecuComp database and COMPUSTAT.

Notes: Generalized Tobit marginal effects are reported with standard errors in parentheses. Two-digit industry and year effects are included in the specification.

[a]The sample size is smaller than it is in Table 1 because of the inclusion of the lag-grants variable.

** Statistically significant at the 5 percent level.

The level of prior grants and the two-year cumulative stock return both have positive and significant effects on current realizations. Higher prior grants of stock options raise both the probability that a firm has positive realizations in the current year (Table 2, column 1) and the expected magnitude of these realizations conditional on the realizations being positive (Table 2, column 2). Holding constant the level of prior grants, we see that current realizations are sharply increasing in the firm's two-year cumulative stock return. Like prior grants, higher stock returns increase both the incidence and magnitude of current realizations. These findings are consistent with the results of previous case studies (Huddart and Lang 1996).

Although larger firms are more likely to experience positive stock option realizations in a year, firm size as measured by employment has no significant impact on the conditional magnitude of the realizations. Similarly, holding constant our other control variables, we observe that the degree of firm leverage has no significant impact on realizations. Stock option realizations show a strong positive relationship with a firm's market-to-book value, reflecting a positive effect of the value on the incidence and magnitude of realizations. Furthermore, higher stock risk raises the likelihood that a firm will experience positive realizations, but it has no further impact on the magnitude.

The Effect of Stock Options on Compensation per Hour

We now assess the overall impact of stock options on aggregate compensation per hour, using our earlier estimates to predict grants and realizations for all COMPUSTAT firms. We use actual firm data on grants and realizations where reported in the ExecuComp sample. For COMPUSTAT firms not in the ExecuComp sample for which we have a complete set of control variables (where we use predicted instead of actual lag stock option grants), we predict stock option realizations using the estimated model in Table 2. For the remaining COMPUSTAT firms for which we have missing data for one or more control variables, we impute their stock option realizations.[22] We aggregate these actual and estimated grants and realizations across all publicly traded firms, and then multiply by the assumed percentage of employee stock options that represent nonqualified stock options.[23] This calculation provides our estimate of the total income generated from nonqualified stock options in that year.

Table 3 shows total compensation for the private, nonfarm business sector, our estimates of total nonqualified stock option grants and realizations, and the growth rates of all three for 1995 through 1999. The findings are presented both on an aggregate and a per-worker basis. The data indicate that in the mid-1990s, stock option grants and realizations amounted to less than 1 percent of total compensation. However, the growth rates of both have significantly exceeded the growth rate of compensation. For example, stock option grants and realizations in 1998 grew by 56 percent and 53 percent, respectively, whereas total compensation grew by 8 percent. Over the five years from 1995 to 1999, stock option realizations per worker more than doubled, from $395 to $1,068.[24]

The rapid rate of increase in the magnitude of employee stock options raises the possibility that they had a significant impact on CPH growth in recent years. This growth can be expressed as a weighted average of the growth in stock option realizations per hour and the growth in other compensation per hour. The weight on the growth in stock option realizations per hour is the share of realizations in that year to total compensation. Despite the small weight given to stock option realizations per hour, their fast growth rate, as seen in Table 3, may imply a significant contribution to compensation growth.

Stock options may be changing the traditional relationship between unemployment rates and pay measures.

Our estimates of the effect of stock options on CPH growth are provided in Table 4. For reference, we also include the annual growth in CPH (column 2). We start by recomputing the growth rate in each year and removing from total compensation an estimate of overall nonqualified stock option realizations in public companies. We do this using two different approaches. First, we perform a simple extrapolation from the ExecuComp sample, which requires no estimation (column 3). For each year, we calculate the average stock option realizations per employee based on all firms in the ExecuComp sample. We then gross this figure up to cover all public firms by assuming that all workers in public firms not in the ExecuComp sample realized this average value of stock options. Our second (and preferred) approach is to use our estimates from Table 2 to predict stock option realizations for COMPUSTAT firms not in the ExecuComp sample (column 4). In both cases, we subtract the implied income derived from

Table 3

Trends in Compensation, Stock Option Grants, and Stock Option Realizations

Year	Compensation (Billions of Dollars)[a]	Stock Option Realizations (Billions of Dollars)[b]	Realizations as a Percentage of Compensation	Stock Option Grants (Billions of Dollars)[c]	Grants as a Percentage of Compensation
			Panel A: Aggregate		
1995	3,488.1	38.6	1.1	26.5	0.8
	(4.6)	(84.2)		(7.3)	
1996	3,656.9	49.8	1.4	39.6	1.1
	(4.8)	(28.8)		(49.6)	
1997	3,911.1	71.6	1.8	55.6	1.4
	(7.0)	(43.8)		(40.3)	
1998	4,214.7	109.3	2.6	86.7	2.1
	(7.8)	(52.7)		(55.9)	
1999	4,489.1	116.0	2.6	110.5	2.5
	(6.5)	(6.1)		(27.5)	
			Panel B: Per Worker[d]		
1995	35,631	395	1.1	271	0.8
	(1.5)	(78.8)		(4.2)	
1996	36,498	497	1.4	395	1.1
	(2.4)	(25.9)		(46.2)	
1997	37,925	694	1.8	539	1.4
	(3.9)	(39.7)		(36.3)	
1998	39,749	1,032	2.6	817	2.1
	(4.8)	(48.6)		(51.7)	
1999	41,333	1,068	2.6	1,018	2.5
	(4.0)	(3.6)		(24.5)	

Source: Authors' calculations, based on data from Standard and Poor's ExecuComp database and COMPUSTAT.

Notes: Stock option realizations and grants are estimates based on ExecuComp and COMPUSTAT data. Percentage changes from the prior year are in parentheses.

[a] Private, nonfarm business sector.
[b] Public companies only—scaled by 82 percent to reflect nonqualified stock options (estimated).
[c] Public companies only—scaled by 50 percent to reflect effective cash value (estimated).
[d] We use the same sample and scaling as we do in the aggregate panel.

nonqualified stock option realizations from total compensation in that year, and we divide by total hours to recompute CPH net of the effect of stock option realizations.[25]

Our calculations reveal that the actual growth in nominal CPH accelerated from around 2 percent in 1995 to 5 percent in 1998, consistent with the labor market tightening that occurred during this period. Notice, however, that if we had removed the contribution of stock option realizations from public companies using our second approach, CPH growth in 1998 would have been 4.3 percent. Thus, stock option realizations appear to have contributed around 0.7 percentage point to CPH growth in 1998. This finding illustrates the sizable impact that a fast-growing segment of compensation can have on overall CPH growth rates, even when that segment still accounts for a small fraction of overall compensation.

Having removed the influence of current stock option realizations from CPH, we now recalculate CPH growth by including the estimated cash value of new employee stock option grants (column 5). This last adjustment yields a CPH measure that should reflect current labor market conditions more accurately. To recalculate CPH, we add the cash value of new employee stock option grants to total compensation less stock option realizations in that year and divide by total hours.

It is now reasonable to ask whether the peculiar way in which stock options enter CPH offers an explanation for the second pay puzzle of the 1990s. To answer this question, we

Table 4
Effect of Employee Stock Options on Compensation per Hour Growth

Year	Actual Growth	Public Companies: Growth less Realizations[a]	Growth less Realizations[b]	Growth less Realizations plus Grants[b]
1995	2.09	1.41	1.61	1.63
			(0.07)	(0.07)
1996	3.07	2.78	2.82	3.14
			(0.07)	(0.07)
1997	3.52	3.12	3.05	3.39
			(0.09)	(0.09)
1998	5.08	3.79	4.30	4.95
			(0.09)	(0.09)
1999	4.64	3.31	4.65	5.05
			(0.09)	(0.08)

Source: Authors' calculations, based on data from Standard and Poor's ExecuComp database and COMPUSTAT.

Notes: Private, nonfarm compensation per hour growth. Monte Carlo standard errors based on 1,000 simulations are in parentheses.

[a] Simple extrapolations based on ExecuComp data.

[b] Based on estimated models presented in Tables 2 and 3.

examine the implied growth rate of CPH in 1998 and 1999 in which we have removed stock option realizations and included new grants. Despite continued labor market tightening in 1999, actual CPH decelerated from its 1998 growth rate, from 5.08 percent in 1998 to 4.64 percent in 1999. However, when we look at our adjusted CPH measure, we find that it continued to accelerate through 1999, from 4.95 percent in 1998 to 5.05 percent in 1999.[26] The drop-off in the pace of actual CPH in 1999 can therefore potentially be explained by the rapid increase in new stock option grants in that year relative to current realizations from prior-year grants.

An implication of this finding is that stock options may be changing the traditional relationship between unemployment rates and pay measures. If firms increasingly use stock options as a substitute for wage and salary increases to attract and retain workers in a tight labor market, the impact of tight labor markets will either be muted in the data (for pay measures such as the employment cost index, which do not reflect stock options), or it will show up with a several-year lag (for pay measures such as CPH, which reflect realizations) because of the vesting requirements for stock options.

The most comparable effort to assess the impact of stock options on aggregate pay measures was made by Lebow et al. (1999). They construct a sample of employee stock option plans for 125 S&P 500 firms from 1994 to 1998. Using the details of the option grants, they calculate modified Black-Scholes values for the new grants in each year. They find that over their sample period, the average stock option grant value per employee grew at a 31 percent annual rate. This result accords well with our data, which indicate an average annual 33 percent growth rate over this period. Assuming that all workers at public companies experienced the same growth rate in stock option grants, the authors calculate that the treatment of stock option grants as compensation on the grant date would have added roughly a quarter percentage point to growth in the employment cost index.

Additional Implications

Our analysis reveals that although employee stock options still represent a small fraction of total compensation in the United States, they have grown rapidly over the past few years. Accordingly, the recent growth in CPH has been significantly affected by the behavior of stock option grants and realizations (Table 4). These findings have several important implications.

If the trend in stock option use continues, CPH growth is likely to be more variable in the future than it has been. As we observed, current stock option realizations depend to a great extent on a firm's recent stock performance. Swings in the equity markets will generate swings in stock option realizations that are likely to exceed the underlying movements in base wage and salary income. This increased volatility suggests that it will be more difficult to discern trend changes in CPH growth. Therefore, an understanding of the effect of stock options on CPH is critical for one to make the correct inference on the underlying pay trends. As such, more accurate and timely data on stock options are clearly needed.

A greater reliance on stock options may also increase overall pay flexibility in the U.S. labor market. Various arguments have been put forward as to why employers are reluctant to impose nominal wage cuts on workers during adverse times (see Lebow et al. [1995]; Groshen and Schweitzer [1996]; Card and Hyslop [1997]; and McLaughlin [1999]). A corollary is that some inflation is good for labor market efficiency because it allows for real wage reductions, even in the absence of nominal wage reductions. Stock options by design build in downward pay flexibility. As noted earlier, the typical nonqualified stock option is issued with a strike price equal to the market price. If the firm does not produce equity gains during the vesting period, the options will remain out-of-the-money and will not be exercised by employees.[27] This added pay flexibility may help to relax any constraints imposed by nominal wage rigidities that exist in the base wage and salary components of pay. Consequently, the labor market may be able to operate efficiently at a lower steady-state rate of inflation.

Furthermore, stock options may strengthen the link between pay and performance. Hall and Liebman (1998) argue that the rising importance of stock options in executive pay has been the primary determinant of the increased sensitivity of executive compensation to firm performance. As stock options filter down the salary ranks, an increasing segment of a firm's salary liability will become linked to firm performance. This restructuring of the wage contract between a firm and its workers therefore may be contributing to the upturn in labor productivity (see Black and Lynch [2000]).

Conclusion

Between 1995 and 1998, actual growth in nominal compensation per hour accelerated from approximately 2 percent to 5 percent. Yet as labor markets continued to tighten in 1999, CPH growth paradoxically slowed. In this article, we have attempted to solve this aggregate wage puzzle by exploring whether changes in pay structure—specifically, the increased use of employee stock options—can account for the behavior of CPH in the late 1990s.

We conclude that the behavior of CPH can be explained largely by the point in time when employee stock options are captured in this measure. When we recalculate CPH growth to reflect the value of current stock options when they are granted—rather than their value when they are realized—we find that our adjusted CPH measure accelerated in each year from 1995 to 1999. This finding suggests that in 1999 there was likely no downturn in CPH growth.

Appendix A: Data Definitions and Descriptive Statistics

Description	Source	Method of Calculation	Mean	Standard Deviation	Minimum	Maximum
Panel A: Grants Regression						
Total grant value (thousands of dollars)	ExecuComp	Grants to top 5 * (100/% of grants to top 5)	31,378	356,183	15	14,200,000
Log total grant value	ExecuComp		8.78	1.59	2.74	16.47
Return on assets, one-year lag	COMPUSTAT	(Operating income before depreciation [13] + interest expense [15]) / (total assets [6])	0.16	0.11	-1.55	0.97
Employment (thousands)	COMPUSTAT	[29]	17.99	49.83	0.01	825.00
Log employment	COMPUSTAT	*ln*([29])	1.49	1.70	-4.96	6.72
Total assets (millions of dollars)	COMPUSTAT	[6]	5,109	21,675	14	668,641
Log total assets	COMPUSTAT	*ln*([6])	6.96	1.63	2.62	13.41
Leverage ratio	COMPUSTAT	Total long-term debt [9]/total assets [6]	0.19	0.16	0.00	1.75
Market value/book value, one-year lag	COMPUSTAT	(Price-close calendar year [24] * shares outstanding [25] + total assets [6] - common equity [60])/ (total assets [6])	2.20	2.08	0.48	45.33
Standard deviation of stock returns, one-year lag	CRSP or Campbell and Lettau	Firm-level (CRSP) data, if available, otherwise industry-level (Campbell and Lettau, *Journal of Finance*, forthcoming) data used	0.02	0.01	0.00	0.09
Number of years since stock first traded publicly	CRSP	Data year minus year stock first traded publicly	13.60	9.85	0.00	30.00
Net operating loss	COMPUSTAT	1 if [52] > 0; 0 if [52] = 0	0.31	0.39	0.00	1.00
Panel B: Probit—Positive Realizations						
Total realized option value (thousands of dollars)	ExecuComp	Realized value for top 5 * (100/% of grants to top 5), two-year lag	6,492	12,693	-2	72,047
Log total realized option value	ExecuComp		5.37	4.02	0.00	11.185
Employment (thousands)	COMPUSTAT	[29]	16.79	45.34	0.005	756.30
Log employment	COMPUSTAT	*ln*([29])	1.53	1.60	-5.30	6.63
Leverage ratio	COMPUSTAT	Total long-term debt [9]/total assets [6]	0.19	0.16	-0.04	1.72
Market value/book value	COMPUSTAT	(Price-close calendar year [24] * shares outstanding [25] + total assets [6] - common equity [60])/(total assets [6])	1.98	1.49	0.49	15.77
Standard deviation of stock returns	CRSP or Campbell and Lettau	Firm-level (CRSP) data, if available, otherwise industry-level (Campbell and Lettau, *Journal of Finance*, forthcoming) data used	0.02	0.01	0.01	0.20
Percentage increase of stock returns over previous two years	CRSP		0.47	1.08	-0.99	18.95
Log total grants, two-year lag	ExecuComp or forecast	ExecuComp data, if available, otherwise forecast used	8.41	1.40	2.74	13.51

Note: COMPUSTAT item numbers are in brackets.

Description	Source	Method of Calculation	Mean	Standard Deviation	Minimum	Maximum
		Panel C: Realized Truncated Regression				
Total realized option value (thousands of dollars)	ExecuComp	Realized value for top 5 * (100/% of grants to top 5), two-year lag	9,602	14,439	0	72,047
Log total realized option value	ExecuComp		7.94	1.86	0.34	11.19
Employment (thousands)	COMPUSTAT	[29]	18.49	47.95	0.04	756.30
Log employment	COMPUSTAT	*ln*([29])	1.63	1.60	-3.30	6.63
Leverage ratio	COMPUSTAT	Total long-term debt [9]/total assets [6]	0.18	0.15	-0.04	1.72
Market value/book value	COMPUSTAT	(Price-close calendar year [24]* shares outstanding [25] + total assets [6] - common equity [60])/(total assets [6])	2.13	1.62	0.56	15.77
Standard deviation of stock returns	CRSP or Campbell and Lettau	Firm-level (CRSP) data, if available, otherwise industry-level (Campbell and Lettau, *Journal of Finance*, forthcoming) data used	0.02	0.01	0.01	0.12
Percentage increase of stock returns over previous two years	CRSP		0.60	1.18	-0.95	18.95
Log total grants, two-year lag	ExecuComp or forecast	ExecuComp data, if available, otherwise forecast used	8.50	1.38	2.74	13.51
Mills ratio	Probit model	ϕ (*Xb* from probit) / Φ (*Xb* from probit)	0.46	0.24	0.06	1.91

Appendix B: Estimating the Level of Stock Option Realizations

We use a generalized Tobit framework to estimate a firm's stock option realizations. Let I^* denote a latent index for the propensity of a firm's workers to realize their vested options in a given year. Let R^* denote desired realizations and R denote actual realizations. We assume that I^* and lnR^* have a continuous distribution, while lnR has a censored distribution.

$$I^*_{it} = Z_{it}\gamma + \varepsilon_{1it},$$

$$lnR^*_{it} = X_{it}\beta + \varepsilon_{2it},$$

$$\begin{bmatrix}\varepsilon_1 \\ \varepsilon_2\end{bmatrix} \tilde{\square} N(0, \Sigma), \ \Sigma = \begin{bmatrix}1 & \sigma_{12} \\ \cdot & \sigma_2^2\end{bmatrix},$$

$$lnR_{it} = lnR^*_{it} \quad \text{if } I^*_{it} > 0$$
$$0 \text{ otherwise.}$$

The probability of a firm having positive stock option realizations in a year is given by $\Phi(Z_{it}\gamma)$. Given the assumed normality of the two error terms, the observed stock option realizations have the following conditional mean:

$$E(lnR_{it} \mid lnR^*_{it} > 0) = X_{it}\beta + \sigma_{12}\frac{\phi(Z_{it}\gamma)}{\Phi(Z_{it}\gamma)},$$

where ϕ and Φ are the standard normal density function and cumulative density functions, respectively. The expected unconditional stock option realizations are given by the probability of observing positive realizations in a year multiplied by the expected conditional magnitude of the realizations.

$$E(lnR_{it} \mid Z_{it}, X_{it}) = \Phi(Z_{it})\left[X_{it}\beta + \sigma_{12}\frac{\phi(Z_{it}\gamma)}{\Phi(Z_{it}\gamma)}\right].$$

We estimate this model in two steps. First, we estimate the γ parameters using a probit model. Using these estimates, we calculate the variable ϕ/Φ for each observation with a positive realization. We then estimate the β parameters by regressing the log positive realizations on our X variables and ϕ/Φ.

Endnotes

1. CPH is the most comprehensive U.S. pay measure. It captures wage and salary income, tips and overtime, paid leave and severance pay, payments in-kind, benefits, bonus and profit-sharing payments, and realizations of stock options.

2. See Farber (1997) and Manski and Straub (2000).

3. See Bell and Neumark (1993), Bell and Kruse (1995), Cohn (1999), Duca (1998), Epstein (1999), and Lebow et al. (1999).

4. We stress, however, that although our conclusion represents an educated assessment of the impact of stock options on the dynamics of CPH, there is a clear need for greater availability of data.

5. See Murphy (1999) for a discussion of the structure of employee stock option plans.

6. A common practice is for a cashless transaction to occur using the services of a third party. The third party makes a short-term loan to the employee to cover the cost of purchasing the exercised options at the strike price. The shares are then immediately sold back to the market and the loan is paid off, with a fee going to the third party.

7. The restrictions are defined in Internal Revenue Code Section 422.

8. For incentive stock options, there is a minimum holding period for the underlying stock that compounds the diversification problem. No similar restriction applies to nonqualified stock options.

9. A Hewitt Associates study of seventy-four plans in 1998 found that 35 percent of the plans used cliff vesting (where all shares vest at the same specified time), with one and three years being the most frequent vesting times; 45 percent used uniform vesting (where shares vest at a uniform rate over the vesting period), with three and four years being the most frequent vesting times; and the remaining 20 percent used either mixed vesting or provided no information (Hewitt Associates LLC 1998). The most common expiration date is ten years after the grant date.

10. This feature of employee stock options makes them a useful tool for reducing employee turnover. Mehran and Yermack (1999) document that the probability of a voluntary departure by a CEO is inversely related to the length of the stock option vesting schedule. They also document that the higher the ratio of deferred compensation to current pay, the less likely a CEO is to leave voluntarily.

11. The median real market value is $8.3 billion for the S&P 500 firms, $1.7 billion for the S&P MidCap firms, and $0.4 billion for the S&P Small Cap firms.

12. In 1994, ExecuComp began recalculating the grant value of a company's new options using a consistent set of assumptions on the interest rate, the implied stock return volatility, and the expected duration of the option. Company handbooks on employee stock option plans typically do not make any distinction between executive and nonexecutive stock option plans. Therefore, we assume that the Black-Scholes value of an option granted to an executive and to a nonexecutive is the same.

13. There is no general agreement as to what caused the popularity of stock options in the 1990s. Murphy (1999) presents a behavioral discussion. Hall and Liebman (2000) examine the role of taxes whereby under Internal Revenue Code Section 162(m), compensation above $1 million is not deductible unless it is performance-based. Of the 1,672 ExecuComp firms in 1998, 1,566 reported paying less than a $1 million salary to their CEOs.

14. This is based on the comparison of COMPUSTAT employment for ExecuComp firms in 1998 with total employment of COMPUSTAT firms in the same year.

15. See U.S. Department of Labor (2000).

16. Our discussion borrows heavily from Yermack (1995).

17. Although volatility always raises the option value of traded stock options, Lambert, Larcker, and Verrecchia (1991) and Kulatilaka and Marcus (1994) show that increased volatility can lower the value of employee stock options, especially for more risk-averse employees.

18. More specifically, the prediction pertains to the relative portion of an executive's compensation that is stock-based. Our dependent variable is the total amount of stock options granted, rather than the ratio of total stock option grants to total compensation.

19. As a robustness check, we also used the firm's "before financing" marginal tax rate (see Graham [1996, 2001]). We found that both net operating loss and marginal tax rates generated the predicted sign and were statistically significant. However, because marginal tax rates were missing for roughly 15 percent of our COMPUSTAT sample, we proceeded using only the net operating loss.

20. The time pattern of exercise for executives may differ from that of other employees for two reasons. First, executives have private information about their firm's prospects that can alter the timing of their exercise decision. Second, footnotes in company proxy statements typically reveal that top executives may exercise their options sooner than the normal vesting schedules permit if certain financial conditions are met. Huddart and Lang (1996) find that the exercise decisions of top management compared with those of other employees are less sensitive to recent stock returns and return volatility.

21. Ideally, we would like to use a weighted average of these ratios based on the typical vesting pattern for employee stock options. However, this is precluded by the short time period covered by the ExecuComp sample.

22. For firms with one or two values missing from our control variables, we impute these values by regressing the variables in question on all other control variables using the estimation sample that has no missing values. We then predict their grants and realizations using these estimated values for the missing right-hand-side variables and actual data for the remaining control variables. For firms with chronic missing data, we leave their grant and realization values missing. We then scale up to a one-digit industry level our estimates to cover all public firms by taking our in-sample average grants and realizations per employee and multiplying the figure by the ratio of total public firm employment in that industry to our in-sample public firm employment in that industry.

23. We assume that 82 percent of employee stock options represent nonqualified stock options, and that this share is constant over our sample period (see Hewitt Associates LLC [1998]). When reporting the value of new nonqualified stock option grants, we scale down first by the 82 percent and then by an additional 50 percent to reflect the likely overestimate of the value by the Black-Scholes methodology (see Huddard and Lang [1996]).

24. Our estimates of the value of stock options per employee are likely to be conservative, given our assumption of no stock option use by privately held firms.

25. To assess the reliability of our estimates, we also report Monte Carlo standard errors. These are computed by simulating draws of new coefficient estimates from the stock option grant and realization estimations, recalculating all results, and repeating this process 1,000 times. The reported standard errors are the standard deviations of the sample distribution of results for each statistic reported in the table.

26. The adjusted CPH growth rate is sensitive to the assumptions we made along the way. For example, if we assume that the cash value of new grants is 60 percent (40 percent) of the Black-Scholes value, the adjusted CPH growth rate in 1999 is 5.13 percent (4.97 percent).

27. However, firms may reprice their employee stock options and/or issue new grants in order to restore incentives. See Carter and Lynch (2000) for Financial Accounting Standards Board reporting of employee stock option repricings.

References

Bell, Linda, and Douglas Kruse. 1995. "Evaluating ESOPs, Profit-Sharing, and Gain Sharing Plans in U.S. Industries: Effects on Worker and Company Performance." Report submitted to the Office of the American Workplace, U.S. Department of Labor. May.

Bell, Linda, and David Neumark. 1993. "Lump-Sum Payments and Profit-Sharing Plans in the Union Sector of the United States Economy." Economic Journal 103 (May): 602-19.

Bizjak, John M., James A. Brickley, and Jeffrey L. Coles. 1993. "Stock-Based Incentive Compensation and Investment Behavior." Journal of Accounting and Economics 16, nos. 1-3: 349-72.

Black, Sandra, and Lisa M. Lynch. 2000. "What's Driving the New Economy: The Benefits of Workplace Innovation." NBER Working Paper no. 7479, January.

Card, David, and Dean Hyslop. 1997. "Does Inflation 'Grease the Wheels of the Labor Market?'" In Christina Romer and David Romer, eds., Reducing Inflation: Motivation and Strategy, 114-21. National Bureau of Economic Research Studies in Business Cycles. Vol. 30. Chicago: University of Chicago Press.

Carter, Mary Ellen, and Luann J. Lynch. 2000. "Does Accounting Affect Economic Behavior? Evidence from Stock Options Repricing." Unpublished paper, Columbia University Accounting Department.

Cohn, Laura. 1999. "The Hidden Costs of Stock Options." Business Week, December 6, p. 44.

Core, John, and Wayne Guay. 1999. "The Use of Equity Grants to Manage Optimal Equity Incentive Levels." Unpublished paper, University of Pennsylvania Wharton School, June.

———. 2000. "Stock Option Plans for Non-Executive Employees." Unpublished paper, University of Pennsylvania Wharton School, January.

Duca, John V. 1998. "The New Labor Paradigm: More Market-Responsive Rules of Work and Pay." Southwest Economy, no 3: 6-8.

Epstein, Gene. 1999. "Sure, Options Pay Represents Funny Money, But Its Earnings Impact Is Less Than Critics Claim." Barron's, October 6.

Farber, Henry S. 1997. "The Changing Face of Job Loss in the United States: 1981-1995." Brookings Papers on Economic Activity, Microeconomics: 55-128.

Graham, John. 1996. "Debt and the Marginal Tax Rate." Journal of Financial Economics 41 (May): 41-73.

———. 2001. "Simulated Tax Rates—Order Form." <http://www.duke.edu/~jgraham/taxform.html>.

Groshen, Erica, and Mark Schweitzer. 1996. "The Effects of Inflation on Wage Adjustments in Firm-Level Data: Grease or Sand?" Unpublished paper, Federal Reserve Bank of New York, January.

Hall, Brian J., and Jeffrey B. Liebman. 1998. "Are CEOs Really Paid Like Bureaucrats?" Quarterly Journal of Economics 113 (August): 653-91.

———. 2000. "The Taxation of Executive Compensation." In James Poterba, ed., Tax Policy and the Economy. Cambridge: MIT Press.

Hewitt Associates LLC. 1998. TCM Data.

Huddart, Steven, and Mark Lang. 1996. "Employee Stock Options Exercises: An Empirical Analysis." Journal of Accounting and Economics 21 (February): 5-43.

John, Teresa A., and Kose John. 1993. "Top-Management Compensation and Capital Structure." Journal of Finance 48 (July): 949-74.

Kulatilaka, Nalin, and Alan J. Marcus. 1994. "Valuing Employee Stock Options." Financial Analysts Journal 50 (November/ December): 46-56.

Lambert, Richard, and David F. Larcker. 1987. "An Analysis of the Use of Accounting and Market Measures of Performance in Executive Compensation Contracts." Journal of Accounting Research 25 (supplement): 85-125.

Lambert, Richard, David F. Larcker, and Robert E. Verrecchia. 1991. "Portfolio Considerations in Valuing Executive Compensation." Journal of Accounting Research 29 (spring): 129-49.

Lebow, David, Louise Sheiner, Larry Slifman, and Martha Starr-McCluer. 1999. "Recent Trends in Compensation Practices." Board of Governors of the Federal Reserve System Finance and Economics Discussion Series no. 32.

References (Continued)

Lebow, David, David Stockton, and William Washer. 1995. "Inflation, Nominal Wage Rigidity, and the Efficiency of Labor Markets." Board of Governors of the Federal Reserve System Finance and Economics Discussion Series no. 94-45, October.

Manski, Charles F., and John D. Straub. 2000. "Worker Perceptions of Job Insecurity in the Mid-1990s: Evidence from the Survey of Economic Expectations." Journal of Human Resources 35, no. 3: 447-79.

McLaughlin, Kenneth J. 1999. "Are Nominal Wage Changes Skewed away from Wage Cuts?" Federal Reserve Bank of St. Louis Review 81 (May/June): 117-32.

Mehran, Hamid. 1992. "Executive Incentive Plans, Corporate Control, and Capital Structure." Journal of Financial and Quantitative Analysis 27 (December): 539-60.

Mehran, Hamid, and Joseph Tracy. 1998. "The Effect of Employee Stock Ownership Plans on Firm Performance." Unpublished paper, Federal Reserve Bank of New York.

Mehran, Hamid, and David Yermack. 1999. "Compensation and Top Management Turnover." Unpublished paper, New York University Stern Business School.

Murphy, Kevin J. 1999. "Executive Compensation." In Orley C. Ashenfelter and David Card, eds., Handbook of Labor Economics. Amsterdam: North-Holland.

Smith, Clifford W., and Ross L. Watts. 1992. "The Investment Opportunity Set and Corporate Financing, Dividend, and Compensation Policies." Journal of Financial Economics 32, no. 3: 263-92.

U.S. Department of Labor. Bureau of Labor Statistics. 2000. "Pilot Survey on the Incidence of Stock Options in Private Industry in 1999." October 11.

Yermack, David. 1995. "Do Corporations Award CEO Stock Options Effectively?" Journal of Financial Economics 39, no. 2: 237-69.

Kenneth N. Kuttner and James J. McAndrews

Personal On-Line Payments

- Personal on-line payment systems—Internet-based systems for making small retail payments—have recently emerged as an alternative to cash, checks, and credit cards.

- All these systems use the web to convey payment information, but they differ in the type of accounts they access: In *proprietary account systems*, funds are transferred between special-purpose accounts maintained by a nonbank provider; in *bank-account-based systems*, funds are transferred between demand deposit accounts at banks.

- Increased acceptance of this payment method will depend on effective risk control and improved settlement arrangements among nonbank providers, a group that currently does not participate in a common clearing system.

- On-line payments are unlikely to have a significant impact on monetary policy, but they do raise some regulatory issues relating to consumer rights and protection.

The rapid growth of e-commerce and the Internet has led to the development of new payment mechanisms capable of tapping the Internet's unique potential for speed and convenience. A recent and especially successful example of such a development is the personal on-line payment: a mechanism that uses web and e-mail technologies to facilitate transfers between individuals.[1]

In a typical transaction of this type, the payer accesses the payment provider's web site to initiate a funds transfer. The payer enters information about the transfer along with payment delivery instructions. Notification of the transfer is sent to the payee by e-mail; confirmation by the payee also occurs via e-mail. The payment provider's computer then transfers the funds.

The first on-line payment systems were created by dot-com start-ups in 1999, and their usefulness quickly became apparent in on-line auctions. These systems grew out of the limitations of retail payment instruments in meeting the needs of auction participants. Most notably, the on-line systems' Internet integration greatly simplified the logistics of making and receiving auction payments. By offering virtually instantaneous funds transfer, the systems made for a much faster payment process than did paper checks, which can take up to five business days to clear. Credit and debit cards, obvious alternatives to checks, have also been unsuitable for most auction sellers because few individuals are equipped to receive payments this way. Moreover, on-line payments are

Kenneth N. Kuttner is an assistant vice president and James J. McAndrews a vice president at the Federal Reserve Bank of New York.
<kenneth.kuttner@ny.frb.org>
<jamie.mcandrews@ny.frb.org>

The authors are grateful to Stephanie Heller for insightful comments and to two anonymous referees for additional suggestions. The views expressed are those of the authors and do not necessarily reflect the position of the Federal Reserve Bank of New York or the Federal Reserve System.

inexpensive compared with credit and debit cards, whose providers typically charge a fee of at least 2 percent of the transaction, with even higher fees for smaller merchants.[2]

Recently, major financial institutions have also begun to offer similar personal on-line payment services—a sign of the systems' increasingly widespread acceptance.[3] Today, there are at least twelve providers. The volume of personal on-line payments, however, is still tiny compared with check volumes. Although comprehensive data are not available for these relatively new systems, a rough estimate is that 500,000 personal on-line payments are made each day and $20 million resides in the accounts of the payment providers. By comparison, check volumes were approximately 186 million payments per day in 1999 and $600 billion was on deposit in domestic commercial bank transaction accounts in 2000.[4]

The distinguishing feature of the personal on-line payment instrument is its use of the Internet for communicating payment information.

Nevertheless, the number of personal on-line payments has grown rapidly over the past three years, and their use in the United States has already surpassed the use of other new electronic methods such as general-purpose "smart cards."[5] Comprehensive industry data again are not available, but one leading payment provider reported 100,000 transactions per day in August 2000 (less than one year after it launched its service) and 200,000 per day by August 2001. Over that same period, the number of users reportedly increased from approximately 3.5 million to 10 million, and the provider estimates that its user base is currently growing at roughly 70 percent per year.

In this article, we examine the personal on-line payment instrument and some of the issues brought to the surface by its development. We sketch the features common to most of the recent instruments of this type as well as draw a distinction between payment instruments based on existing bank accounts and those provided by nonbanks. The problems created by the proliferation of different systems and the importance of interoperability are also addressed, as are issues related to risk management and regulatory and monetary policy. We conclude with some general observations on these unique instruments.

Types of Personal On-Line Payment Systems

The distinguishing feature of the personal on-line payment instrument is its use of the Internet for communicating payment information (Box 1). In fact, these systems are the first to successfully exploit the Internet for that purpose. Typically, payments are initiated from the payment provider's secure web site, with notification taking place via e-mail. This arrangement cleverly utilizes the increasingly ubiquitous electronic address and delivery system to alert a payee that funds have been sent or to request funds from a payer.

How Do the Systems Work?

The first step in a typical transaction is initiation: the payer accesses the payment provider's web site, using a secure, encrypted connection, where he enters the amount of the funds transfer and the e-mail address of the recipient. In the notification step, the provider's computer sends a message to the recipient containing a hyperlink to the provider's web site. Confirmation takes place when the recipient clicks on this link, establishes a secure connection to the provider's server, and confirms the funds transfer.

Although the mechanics of the transaction are similar, the systems differ according to the type of accounts from which the funds are drawn and the payment networks used for completing the transaction. Personal on-line payment systems introduced thus far generally fall into one of two categories: those based on proprietary accounts held at the provider itself and those based on bank accounts. Bank-based systems can be classified further according to whether they use Automated Clearing House (ACH) or automated teller machine (ATM)/point-of-sale (POS) debit card payment networks.

Personal on-line payment systems introduced thus far generally fall into one of two categories: those based on proprietary accounts held at the provider itself and those based on bank accounts.

Proprietary Account Systems

The first systems introduced were based on proprietary, nonbank accounts. In these systems, values are transferred between special-purpose accounts created and maintained by a nonbank provider. Deposits to the account can be made using a credit card, directly from a bank account via the ACH network, or by paper check. An important advantage of these systems is the extremely quick and simple process of completing intraprovider payments: the payment is made through a book-entry transfer and occurs almost immediately after the receiver acknowledges receipt of the e-mail.

A typical person-to-person payment transaction is presented in Exhibit 1. The payer begins by transferring ("downloading") funds from an existing bank or credit card account to his account at the payment provider. The payer then

Box 1

What Is a Payment Instrument?

A payment is a transfer of monetary value from one person to another. A payment *system* is the mechanism—the rules, institutions, people, markets, and agreements—that make the exchange of payments possible. In general, three elements are required to accomplish this task on a widespread scale: a secure communications system, a set of accounts in which the value to be transferred is stored, and a method of moving value from one account to another. The last element is sometimes called the clearing and settlement system. A payment *instrument* consists of the instructions to transfer value bundled together with the communications system. A payment instrument may use a unique clearing and settlement system or one that is shared among many payment instruments.

Consider a check. Checks are nothing more than written instructions, delivered by hand or by mail, directing the payer's bank to transfer account balances from the check writer to the payee. The payee's bank utilizes the system for clearing check payments to have funds transferred from the check writer's bank to it, typically at a collecting bank at which they both hold accounts. An analogous arrangement characterizes credit card transactions. For payments processed electronically, the card, together with the terminal, creates instructions communicated (in an encrypted format) over telephone lines to transfer money from a line of credit of the cardholder to the payee, again using a clearing and settlement system to transfer funds between the banks involved.

Cash is unique as a payment instrument in that it *self-clears*. It represents value (a liability of the central bank) that is not in an account, but is instead a circulating liability. This feature of cash is supported, at least in part, by its role as legal tender—that is, cash discharges a debt by force of law. As such, the value is transferred at the same time the (hard-to-counterfeit) cash is exchanged. In effect, the communications system for cash, which is hand-to-hand transfer, also provides its clearing and settlement mechanism.

Wholesale payment systems, such as the Federal Reserve System's Fedwire funds transfer service, work in a similar fashion to checks and credit cards. Instructions to transfer funds flow through a communications network operated over telephone lines to the Federal Reserve and its participant banks. The Federal Reserve then deducts funds from the account of the sending bank, credits them to the account of the receiving bank, and notifies both banks of the completion of the transaction. Because of the large amounts that are often involved, wholesale systems typically restrict participation to banks, although banks can offer their customers the ability to use these systems indirectly, while retail systems are intended for widespread use by households and firms. In addition, most wholesale payment systems today offer final settlement of the funds transfer between the two banks on the same day—if not in the same minute—as the instruction is entered, while retail systems typically offer final settlement with a delay of at least one day.[a]

Payment instruments can be differentiated according to whether they provide distinct means of conveying the instructions to transfer balances between the payer and the payee. Different instruments may be used to transfer value into or out of the same account: checks and Automated Clearing House debits, for example. By the same token, it is also possible for the different instruments to utilize the same underlying clearing and settlement system. In fact, many personal on-line payment instruments do just that.

[a]Settlement becomes final when the transfer of funds has occurred and is irrevocable, even, for example, if the payer's bank fails.

initiates the transfer on the provider's web site, the recipient is notified via e-mail, and the transfer is confirmed. Once the process is complete, the provider's computer transfers the value between the two users' accounts. The recipient can leave the funds in the account for future use, or she may opt to move them to a traditional bank or credit card account.

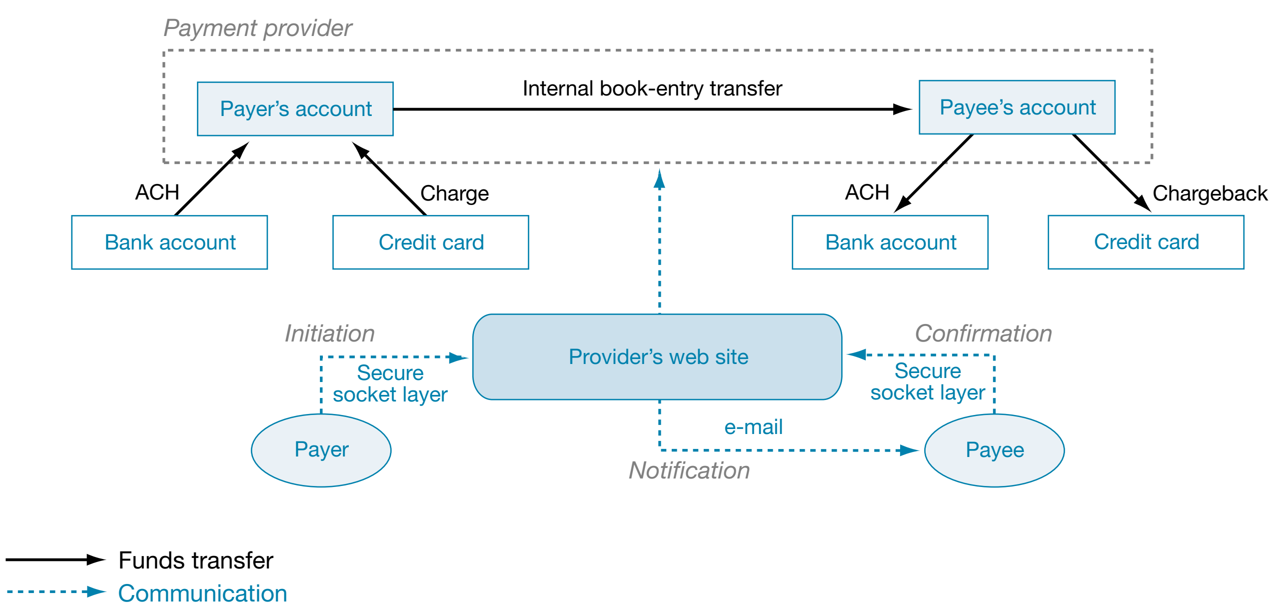
Two features unique to proprietary account systems are worth noting. First, payments to payees *not* signed up with the same provider either require the payee to establish an account or the provider to use a conventional payment instrument to effect the funds transfer. A payment destined for a bank demand deposit account (DDA) would utilize the ACH network or a paper check. Alternatively, the payment could be completed via a "chargeback" to the payee's credit card. Second, payers who do not wish to maintain a balance with the provider usually can charge payments to a credit card—effectively a just-in-time transfer of value into the account. A key factor in determining whether users decide to maintain positive balances, or opt instead to upload and download funds as needed, is the frequency with which they expect to make payments; this, in turn, will depend on whether on-line payments become widely accepted and the degree of interoperability between competing systems. Broader acceptance and greater interoperability will tend to increase the usage of on-line payments and thus lead to larger average balances maintained in providers' accounts.

Thus far, these proprietary on-line payment services have remained free for consumer transactions, although the transfer of funds to or from the provider sometimes incurs a fee. In an apparent effort to encourage the use of credit cards for such services, credit card companies usually treat the download of value into these accounts as a sale, rather than as a cash advance, which allows the user to avoid interest and cash-advance charges. The provider absorbs the "interchange fee" associated with the transaction, although at least one provider charges a small fee for credit card downloads and ATM withdrawals.[6] Businesses and high-volume individuals typically pay fees for receiving funds and transferring the funds into DDA accounts. The sum of these two fees is approximately the same as the credit card providers' discount fee, making the cost competitive with traditional credit cards.

One drawback of e-mail-based on-line payments is that they are rather cumbersome for person-to-business payments. Most businesses prefer to have funds transferred automatically to an existing account, rather than receiving an e-mail notification and manually confirming each transfer. This limitation has led to a variation on the basic personal on-line payment scheme—one that combines a proprietary account with a "virtual" signature-based debit card. In such a system, the account holder downloads value to his account in the usual way, but initiation takes place on the merchant's web site rather than on the provider's. The transfer of value takes place over a debit

Exhibit 1

Person-to-Person On-Line Payment: Proprietary Account System



card network (either MasterCard's MasterMoney or Visa's VisaCheck, but currently not over a PIN-based debit card system) and settles the next day, just like any other debit card transaction. A major attraction of such a hybrid system is that it leverages the existing debit card network, so it is automatically accepted by the millions of businesses already set up to take debit cards. In addition, the user can employ the debit card in the conventional way to withdraw cash from ATMs and make other purchases at the point of sale.

Bank-Account-Based Systems

Although systems based on proprietary accounts were the first to appear, a number of providers—typically banks—more recently have developed systems that obviate the need to establish a special-purpose transaction account. In these systems, the web and e-mail communications links are similar, but the systems allow a payer to transfer funds directly from his account to that of the payee, even when the payee's account is at a different bank. As in the proprietary-account-based systems, payments from credit cards are also possible, as are payments to credit card accounts via a chargeback transaction. (However, not all providers treat a credit card payment as a purchase rather than a cash advance.)

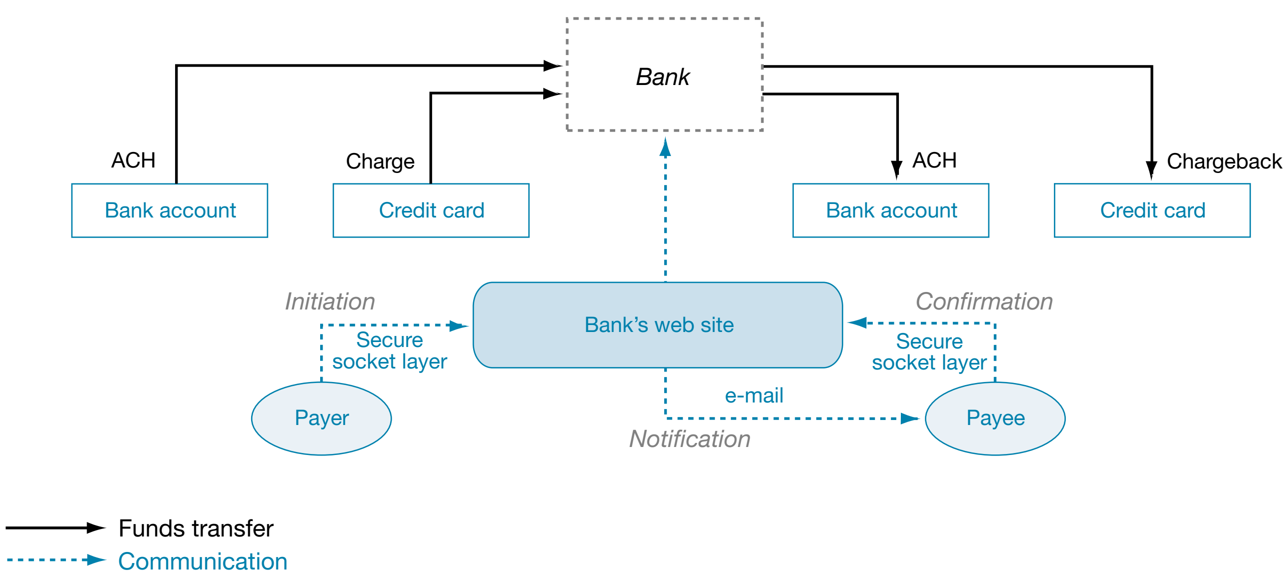
Exhibit 2 depicts a simple transfer of funds in a DDA-based system. The initiation, notification, and confirmation steps are essentially the same as those in the proprietary account system. The main difference is the source of funds and how value is transferred. Here, the accounts accessed are demand deposit accounts at banks, rather than proprietary funds transfer accounts set up by the payment provider. Essentially, the arrangement provides yet another way to access a bank account, supplementing the check, point-of-sale debit card, telephone-based automated account system, and automated teller machine mechanisms.

A further distinction can be made between those bank-based systems that use the ACH network to transfer funds and those that use the ATM/POS network. The ACH network is relatively slow, typically requiring one to two days for transaction authorization and settlement. By contrast, the ATM/POS network provides a real-time authorization and guarantee to the payee that funds are available, and settlement is usually completed the next day. ATM/POS-based transactions currently require the payee to give her ATM/POS debit card number to the payer, however, which adds a layer of complexity, as well as a security concern, not present in systems based on e-mail addresses.

Furthermore, these bank-account-based systems are typically subject to slightly higher fees than the proprietary account systems. These amounts can either be a fixed charge or a fee based on the amount of the transaction. In some systems, the payer bears the cost; in others, the payee incurs it.

Exhibit 2

Person-to-Person On-Line Payment: Bank-Account-Based System



Proliferation and Integration of Systems

The number of providers of personal on-line payments has grown quickly in the past few years, and there are currently at least a dozen. In what has become a familiar pattern for on-line services, it appears that several firms have entered the business with the intention of quickly gaining market share. This pattern stems from the fact that payment instruments display what are known as positive network effects (McAndrews 1997), which accrue when an increase in the number of users of a good makes the individual user better off. In this case, more widespread acceptance of a certain provider's system makes that provider's services more useful and convenient for an individual user, by virtue of his ability to send payments to and receive payments from a wider circle of people. Product differentiation can work against the network effects, however, as differences between products can be valuable to users with different tastes or needs. If providers cater to customers with distinct tastes, the benefits of product differentiation can outweigh the gains from having a single dominant provider.

If no single firm establishes a dominant position, the fragmentation of the market among competing providers could limit the usefulness and convenience of any provider's product. One way to mitigate that problem would be for different providers to employ some form of interoperability standard, as long as interoperability did not itself prevent product differentiation.[7]

Clearance/Settlement and Interoperability

What does it mean to make payment instruments interoperable? Interoperability allows an account holder at one provider to make a payment to or receive a payment from an account holder at another provider.[8] This can be accomplished if both providers participate in a common clearing and settlement system, the purposes of which are to account for transactions, transfer payment messages between the providers, and arrange for the transfer of settlement balances (such as balances at a correspondent bank or the Federal Reserve) among providers.

Familiar check-based payments are interoperable in this sense: existing clearing and settlement systems allow a person who holds an account at one bank to pay an account holder at another bank by writing him a check. The check then travels between the payee and the two banks by way of the clearing and settlement system, and one bank transfers funds to the other at a third bank at which they both hold deposits.

The systems of the nonbank personal on-line payment providers currently are not interoperable in the same way: no clearing and settlement system exists that would allow an account holder on PayPal, for example, to send funds directly to an account holder on ecount.[9] To effect such a transfer, either the payer or the payee must first register as an account holder with the other provider and then complete the transaction using other payment instruments. Suppose, for instance, that the payee signed up with the payer's provider. The recipient would then have to request that the funds be transferred to her credit card, wait for the funds to clear, and then use the credit card to add funds to her original account. Thus, the systems provide only *indirect* interoperability through their use of other payment instruments, such as credit cards and checks.

The lack of interoperability imposes obvious costs on users, such as the inconvenience of maintaining accounts at multiple providers, not to mention any fees associated with transferring funds between the various accounts. Yet despite these costs, providers may still prefer not to make their systems interoperable. Notwithstanding the expense involved in creating and managing a clearing and settlement system, interoperability can work to make the products more closely substitutable and thereby increase the competition between the products. The requirement to open an account in order to receive funds, for example, has been a key element in some providers' "viral" marketing schemes. Although costly, the lack of interoperability is not necessarily inefficient, however, especially in such a dynamic, rapidly evolving industry. The gains from imposing full interoperability through governmental or industry action would need to be weighed against the gains from product differentiation and further innovation.

> *Payment instruments display . . . positive network effects, which accrue when an increase in the number of users of a good makes the individual user better off.*

In Canada, the interoperability issue has been resolved by mutual agreement among that country's five leading banks. The agreed-upon system uses Canada's single clearing and settlement system for interbank payments to transfer funds between banks (for those recipients who do not wish to receive funds on their credit cards). Two factors help explain Canada's rapid, unanimous adoption of a single system. One is Canada's relatively concentrated banking industry (most demand deposits reside at one of the five largest banks). The second is the country's unified payment infrastructure, which uses a single clearing and settlement system for all interbank transfers, large and small: check, ATM, ACH, and POS.

The fully interoperable Canadian system provides an interesting juxtaposition to the more fragmented, less bank-oriented American system. Given the competitive industry dynamics in the United States, it is unlikely that the different payment providers will agree upon the adoption of a new clearing and settlement system in the near future. Instead, the providers will likely continue to rely on the interoperability of each of their systems with the clearing and settlement systems of banks and credit card associations.

There are, however, a number of steps that nonbank on-line payment providers could take to improve interoperability, short of adopting a common clearing and settlement system. For instance, the providers could each hold an account in a common bank, which would effectively serve as the system's settlement agent: when transfers are completed in the common bank, settlement would be complete. Another step would be for providers to employ a bank to make interbank funds transfers on their behalf (or even establish a bank for this purpose). Finally, providers could set up a clearing house arrangement for netting and settling payments.

Policy Issues

Risk

Providers of on-line payment instruments are concerned about the risks of fraud, operational failure, and other liquidity and credit risks because their success depends on maintaining a system that is useful to customers and protects the provider from fraudulent withdrawal of funds from the system. Therefore, it is important to examine the risk control measures employed by these new systems to combat risk.

Fraud is perhaps the most immediate threat faced by on-line payment providers. To address this risk, all the systems register and communicate credit card information using a secure socket layer—an encrypted connection to the provider's web site. The payer's information is retained by the provider, reducing the need for repeated transmission over the Internet. Another risk control is a limit on the size of payments that can be made. Some providers, for example, limit transfers to very small amounts until the user's identity and address are verified by conventional mail.

Risk is also posed by the extensive use of e-mail. The systems use this medium for various purposes: e-mail serves as a means of communication, the e-mail address acts as an addressing or locating system, and one's e-mail response to a receipt of payment is used, in part, as a means of identifying the payment recipient. A single e-mail account shared by several people naturally will diminish the effectiveness of e-mail as an identifier and a means to communicate to only one person. As a result, additional means to identify the recipient become necessary. Increasing the number of hurdles a user must overcome to transfer value may lower system risk, but at the cost of reducing system convenience.

> *The success [of personal on-line payments] depends on maintaining a system that is useful to customers and protects the provider from fraudulent withdrawal of funds from the system.*

It is worth noting that the leading personal on-line payment provider grew out of an encryption firm, which indicates that the sponsors recognize the importance of preventing counterfeit and fraudulent claims from being entered against the company. One company official stated that successful providers will have to supply world-class fraud prevention and detection systems to manage this type of risk. If these systems should mature and create a more universal, interoperable system, then the operational risks will loom larger simply because of the larger values involved. In the meantime, it is safe to say that the existing systems are already under intense scrutiny by security experts (as well as hackers) for any possible weaknesses.

Like traditional financial intermediaries, on-line payment providers also face a certain amount of credit and liquidity risk. So far, this risk has been relatively modest: the dollar amounts involved have been too small to create significant risk for the financial system. In addition, nonbank providers generally maintain the assets in money market funds or at banks, all but eliminating credit and liquidity risk. Therefore, as long as providers continue to keep their funds in short-term, high-quality assets, credit and liquidity risk will not be a major issue.

Regulatory Treatment of Payments

One important issue that could affect the acceptance of on-line payments concerns the rights of consumers when using this payment method. The use of credit and debit cards is governed by a well-established set of legal rights, in addition to any contractual terms agreed to by the card issuer and consumer. In some cases, federal regulations grant consumers a certain amount of protection against fraudulent use of their cards as well as certain rights in case of errors made by the payment provider, including certain rights to resolve errors. Furthermore, consumers' potential losses are limited under the regulations that govern those card payments.[10]

In contrast, most personal on-line payments (in particular, those based on proprietary accounts) do not involve a credit or debit card, and therefore the consumer may not enjoy the same set of legal rights that he would in a credit or debit card payment. These rights are governed instead solely by the private contractual terms set out by the providers in the user agreement. It is not clear whether consumers are aware of this distinction, particularly as many of them fund their on-line accounts with a credit card in the first place.[11]

Regulatory Treatment of Payment Providers

Some personal on-line payment providers are banks and some are not, and this distinction gives rise to differences in regulatory treatment. Bank providers, for example, are required to hold a certain share (3 or 10 percent, depending on the level of deposits) as non-interest-bearing reserves, while nonbank providers currently have no such requirement. In addition, unlike nonbank providers, banks are required to hold a minimum level of capital. Banks are also subject to reporting requirements and periodic examination by supervisory authorities such as the Comptroller of the Currency, the Federal Reserve, the Federal Deposit Insurance Corporation, and state banking agencies. Finally, banks can avail themselves of deposit insurance for account balances up to $100,000, while nonbank providers cannot offer this protection.

Because nonbank providers of personal on-line payments typically have chosen to invest in low-risk assets, the providers resemble "narrow banks"—institutions that hold only risk-free, liquid assets, and by doing so avoid the threat of bank runs.[12] Because of this feature, narrow banking is sometimes proposed as a way to render deposit insurance unnecessary. (Nonbank payment providers are not *required* to disclose this information, though.) Consequently, there is probably little demand for traditional deposit insurance. Fraud, however, is a major concern. In light of this concern, some on-line payment providers have offered private insurance against fraudulent use of their customers' accounts, to enhance the attractiveness of their service. (This differs from deposit insurance, however, which insures against bank insolvency.)

These issues raise the question of whether nonbank personal on-line payment providers are in effect banks. The answer depends on the definition of "bank." If a bank is an institution that "takes deposits and makes loans," the answer would be no, as these providers typically invest in money market assets, rather than loans.[13] This is not the only definition of a bank, however. An alternative definition, codified in the Glass-Steagall Act, focuses on the role of banks as deposit takers. The Act precludes any institution other than a state-licensed money transmitter or a state or national bank from engaging in "the business of receiving deposits subject to check or to repayment upon presentation of a passbook, certificate of deposit, or other evidence of debt, or upon request of the depositor."[14] From an economic perspective, as receivers of funds subject to withdrawal or transfer upon the instruction of customers, nonbank on-line payment providers might be deemed to fit this definition.[15] Alternatively, certain nonbank providers of arguably similar services—for instance, money transmitters such as Western Union and traveler's check firms such as American Express—are legally recognized and are licensed in several states to provide these services.

Some personal on-line payment providers are banks and some are not, and this distinction gives rise to differences in regulatory treatment.

The resemblance of personal on-line payment providers to narrow banks also raises the issue of the complementarity between lending and deposit taking emphasized in various theories of banking. Some recent theories—such as those of Diamond and Rajan (2001a, b), Kashyap, Rajan, and Stein (forthcoming), and McAndrews and Roberts (1999)—suggest that the provision of transaction deposits naturally lends itself to wider banking activity, such as lending. In those theories, the provision of transaction deposits creates a form of liquidity that can be utilized to make loans more cheaply than those offered by other firms. If those theories are correct, and apply to the personal on-line payment providers, then the providers might be transformed, over time, into more bank-like firms to take advantage of this economy. In contrast, the transmitters of small-value wire transfers and traveler's checks, although similar to personal on-line payment providers, have not transformed themselves into lenders, as these theories might imply.

These theories suggest that firms that provide only payment services and maintain funds in low-risk and low-return assets may have a difficult time recovering all the costs of their activities. Again, the evidence provided by the long-term viability of wire transfer and traveler's check firms suggests that some payment providers can earn normal profits without engaging in full-fledged banking. This evidence is supported by alternative theories of banking that maintain that the complementarity between deposit taking and lending is weak, and that the two activities can be separated without sacrificing efficiencies.

Monetary Implications

The development of new electronic payment mechanisms raises interesting questions for monetary policy. Some observers have even expressed concern that introducing what is in effect a private currency may undermine the Federal Reserve's control of monetary policy. Such concern is unwarranted, however, at least for the personal on-line payment systems described here. Although the widespread adoption of on-line payments may alter some familiar economic relationships or call for technical changes in the definitions of the monetary aggregates, on-line payments in their current form will not threaten the Federal Reserve's ability to conduct monetary policy in the foreseeable future.[16]

Perhaps because these systems all rely on the Internet, it is easy to overestimate their novelty: from a monetary perspective, personal on-line payments are in fact more evolutionary than revolutionary. For years, it has been possible to transfer balances electronically between banks, and the recently introduced bank-based on-line payment instruments do little more than make that technology more readily available to the consumer.[17] In DDA-based systems, web and e-mail technologies therefore merely replace paper checks as payment instructions and communications links.

The growth of personal on-line payments at *nonbank* providers could have some monetary effects, but these would be of a minor, technical nature. Data on balances held at nonbank providers are not collected, nor are they included in the monetary aggregates (such as M1) as currently defined. This would make the aggregates an imperfect gauge of the reservoir of "liquidity" available for spending, and could further reduce the aggregates' information content. If this were a concern, the impact could easily be remedied by extending the aggregates' definition to include balances held at nonbank providers.[18] In any case, prior waves of financial innovation have already diminished the information content of the monetary aggregates to the point where they have lost their status as policy targets.[19] A related, and equally inconsequential, effect would be any change in the so-called money multiplier brought about by changes in the reserve-to-deposit or currency-to-deposit ratios, similar to the change brought about by the adoption of "sweep" accounts in recent years.

From a monetary perspective, personal on-line payments are in fact more evolutionary than revolutionary.

These technical factors may have some effect on the monetary aggregates, but they are unlikely to have a discernible impact on U.S. monetary policy or its efficacy. Although the textbook description of monetary policy involves the control of the money supply,[20] current Federal Reserve policy focuses on an interest rate target, adjusting open market operations in light of changes in money demand. Current policy would thereby insulate interest rates from any shift in consumers' preferences from reservable bank deposits, for example, to nonreservable balances at on-line payment providers. Only if the volume of bank liabilities *per se* mattered, would such changes affect monetary transmission.[21]

Looking ahead into the more distant future, some observers have imagined a world in which virtually *all* payments will have gone on-line, drastically reducing banks' demand for reserve balances.[22] Leaving aside questions of plausibility, such an extreme situation would undoubtedly present new and unforeseen challenges to monetary policy, and it is reasonable to ask whether policy would remain effective. A complete analysis of that possibility is well beyond the scope of this article. However, it is important to emphasize that all of the on-line payment instruments described here retain the U.S. dollar as the unit of account, and as long as settlement ultimately takes place in dollar-denominated Federal Reserve balances, the Fed will retain leverage over short-term interest rates.[23]

Indeed, payment systems based on nondollar standards have appeared from time to time, and some of these systems make use of the Internet for initiating and confirming payments (Box 2). If one of these systems were to gain widespread acceptance, it could, in theory, pose a more fundamental challenge for monetary policy. However, use of these nondollar systems has remained very limited, and given their costs and limitations, it is hard to imagine that any one of them could ever displace the dollar as the dominant unit of account in the foreseeable future.

Box 2

New Parallel Money

The new payment instruments described in this article are based on the U.S. dollar. Accordingly, although they replace paper currency or checks as a payment instrument, the dollar retains its basic functions as a unit of account and a store of value. The link to the dollar is maintained by the guarantee of convertibility: users may convert balances at par into cash, bank account balances, or credit card balances at any time.

However, not all payment instruments are based on the dollar. Some actually replace the dollar with a proprietary or commodity-based standard and offer only limited convertibility into dollars. We describe below three examples of this "parallel money" and trace its origins to older, nonelectronic systems.

E-Barter

One form of parallel money has been developed for use in on-line barter transactions, or "e-barter," which has grown in popularity among businesses. Typically, these arrangements offer participating businesses a way to trade their excess inventories for needed supplies.[a] Rather than posting prices in dollars, participants quote prices in the sponsoring firm's proprietary unit of account. In most systems, there is a stated notional conversion rate between this unit of account and a dollar (usually one-for-one), and in some cases it is possible to purchase e-barter credits with dollars. However, the provider is generally not committed to redeem these e-barter credits for dollars; consequently, the purchasing power of the credits could, in principle, deviate from dollar-denominated prices.

According to the providers, one advantage of e-barter is that it allows firms to economize on their use of cash. It is not entirely clear, however, how this translates into a real benefit to firms. For instance, if a firm has $1,000 worth of widgets to sell, it is hard to see the advantage in trading those widgets for $1,000 worth of barter credits, rather than cash. Tax avoidance is presumably not a factor, as e-barter providers report to the IRS the dollar-equivalent proceeds of barter transactions. Instead, barter arrangements may facilitate price discrimination—that is, one can sell the inventory at two prices to two types of customers, distinguished by their willingness to accept barter.

Although monetary payment systems have almost completely displaced barter in modern economies, barter systems have occasionally cropped up in the guise of alternative currencies. Perhaps the best-known recent example is Ithaca Hours, which have circulated in the Ithaca, New York, region since 1993.[b] The unit of account in this system is the Hour, which is supposed to reflect the labor used in providing the good or service. Ithaca Hours may be purchased at a fixed exchange rate of $10 for one Hour, but like the e-barter systems, Hours are generally not convertible back into cash.

E-Loyalty Points

Another money-like system involves the "points" some merchants offer their customers in return for the purchase of their products. Accumulated points can then be redeemed for prizes, or for more products of that company or a set of participating partners. The idea is to build brand loyalty by giving customers an incentive for repeat purchases. The ubiquitous frequent-flyer miles are the best-known example of such a system. The idea has recently been extended by e-commerce providers, some of whom have begun to offer "e-loyalty" points for on-line purchases.[c] In fact, the S&H Green Stamps trading stamp program, popular in the 1960s and 1970s, has recently been reincarnated as on-line "GreenPoints."

Although these e-loyalty points resemble money in the way that they define a unit of account, they lack many of the essential features of full-fledged parallel money. Most conspicuously, the points are generally nontransferable, which prevents them from circulating as a medium of exchange.[d] No points are convertible directly into dollars.

New Parallel Money (Continued)

On-Line Commodity-Based Moneys

Perhaps the most radical direction taken by parallel money arrangements combines electronic book-entry transactions with a unit of account and store of value based on a precious metal, such as gold.[e] In this system, account balances represent physical quantities of gold and transactions are ultimately denominated in those terms. Balances can be converted into cash at the prevailing market price.

This arrangement represents the most complete parallel money system, encompassing all the traditional functions of money. Except for the ability to initiate transfers electronically, the system is indistinguishable from the one developed by medieval goldsmiths, in which paper claims on physical stocks of gold circulated in lieu of the metal itself. This arrangement was displaced throughout Europe in the 1400s by the fractional reserve banking system we have today. The main reason for its demise was its high cost: gold earns no interest, yet requires tight security to prevent theft. The same high costs will probably prevent on-line commodity-based moneys from gaining widespread acceptance, despite the advantages offered by electronic access.

---

[a]Examples of e-barter providers are Bigvine, uBarter, and BarterTrust. For an overview, see Lorek (2000).

[b]See <http://www.lightlink.com/ithacahours>. Other local currencies are described in Graham (1996). A classic reference on barter systems is Radford (1945).

[c]Examples include beenz and Yahoo! Points.

[d]The "dollars" issued by the Canadian Tire retailer *were* transferable, however, and enjoyed some limited circulation as a medium of exchange.

[e]Examples of such systems are e-gold, IceGold, and GoldMoney; also see Ballve (2001).

Conclusion

Despite their relatively short track record, personal on-line payment systems have already proven to be especially useful for accommodating small payments made in e-commerce. These systems are also providing a convenient alternative to paper checks, money orders, and cash for a variety of transactions, and they may replace credit cards for some small-scale retail commerce.

Accordingly, we can make a number of observations relevant to the evolution of electronic payments. For example, personal on-line payments were developed in response to a specific market demand, and this fact likely accounts for much of their success. Persons selling items in on-line auctions could not accept credit cards—the usual way of making consumer-to-business payments on-line—and checks represented a payment option that was neither integrated with on-line auctions nor particularly fast or safe. This attention to market demand allowed personal on-line payment providers to put a working payment instrument into the market quickly, and significant volume ensued. In contrast, some electronic stored-value payment instruments, offered by smart-card providers, have been in development for years, and few real market transactions have yet to be processed on the providers' systems.

Furthermore, personal on-line payment instruments use the systems of the credit card associations, the Automated Clearing House, and the electronic fund transfer networks to clear and settle payments. This relationship between the new payment instrument and the existing payment systems is therefore complementary. And while personal on-line payments will probably never replace conventional payment instruments, they may actually increase the number of transactions flowing through the credit card systems (a form of "clicks-and-bricks" synergy). As a result, by utilizing existing systems for clearing and settlement, personal on-line payment providers have been able to succeed where others have failed. The alternative—building a new clearing and settlement network from scratch—would have required large investments and would have offered users of existing systems a less familiar way to make payments.

A final observation concerns these new systems' use of increasingly popular e-mail. The use of e-mail was innovative, and in retrospect it represented a natural approach to delivering payment information on-line. A lesson here is that the use of e-mail to deliver information, while relying on existing secure clearing and settlement systems to transfer value, may ultimately be more practical than more ambitious schemes to transmit digitized stored value directly.

Going forward, we note that other novel uses for these payment systems are in the wings. For instance, most on-line payment providers have plans to offer a capability for making purchases from mobile wireless devices such as cell phones. Another plan is to allow consumers to make anonymous purchases by concealing their identity from merchants. A third goal is to enable small transfers to be made between individuals. Some providers have even marketed their products as a way to fund spending by teenagers and college students while avoiding the risk of overdrafts and overrun credit limits. Finally, another use for these systems involves the delivery of on-line rebates and gift certificates, where integration with the merchant's web site confers a unique advantage over paper checks.

Our review of these innovative systems suggests that although their monetary effects are not likely to be great, their regulatory treatment is likely to evolve. Moreover, the status of various issues relating to personal on-line payments remains open to discussion. These issues include the incorporation of payment transactions in the monetary aggregates, consumer protection rights, the regulatory regime to which the providers are subject, the insurability of deposits in the providers' on-line accounts, and the reservability of the deposits. The development of personal on-line payments therefore compels us to revisit two central questions of economics and law: What is money? and What is a bank?

Endnotes

1. Providers of this type of payment service include PayPal, ecount, Citigroup's c2it, BillPoint, and CertaPay in Canada.

2. See Chakravorti and Shah (2001) for information on credit card fees.

3. However, on-line auctions still represent a major source of the demand for on-line payments: eBay alone reports that on any given day it hosts millions of on-line auctions, with more than 600,000 new items joining the "for sale" list every twenty-four hours. See <http://pages.ebay.com/community/aboutebay/overview/benchmarks.html>.

4. The 1999 check volume data, the most recent available, are from the Bank for International Settlements (2001, p. 105). The level of deposits is from the Board of Governors of the Federal Reserve System (2001, p. A15).

5. General-purpose smart cards differ from "closed-system" smart cards—which are widely used in some transportation, building access, document reproduction, and laundry facilities—in that the general-purpose cards are meant to be accepted and used in a wide range of businesses. Closed-system cards are intended primarily to serve a single use.

6. An interchange fee is a fee charged by the credit-card-issuing bank to the merchant's bank. Merchants pay a "discount fee" that typically is equal to the interchange fee plus a small markup.

7. The Visa and MasterCard credit card systems originally required that their members not issue the other brand of credit card, and if a member issued one brand it could not accept the merchant receipts of the other brand. The two systems later dropped these restrictions, essentially making themselves interoperable. The Discover credit card system and the American Express charge card system operate independently and are not interoperable (according to the definition used in this article) with each other or with Visa and MasterCard.

8. This definition of interoperability focuses on the use of a common clearance and settlement system for the instrument, rather than on technical aspects of interoperability. It is useful to point out that a person cannot pay another person with a check by increasing the line of credit on the recipient's credit card. Instead, the check typically must clear and settle either for a cash payment or for deposit into the recipient's bank account. In that sense, credit cards and checks are not interoperable according to this definition, even though both methods offer a widely used payment system.

9. Interoperability is less of an issue for bank-account-based systems, which all rely on the bank clearing and settlement system for funds transfers.

10. A number of federal regulations affect the rights and responsibilities of parties using credit cards, debit cards, checks, or deposit accounts. For example, Regulation E establishes the rights, liabilities, and responsibilities of parties in electronic funds transfers and protects consumers when they use such systems. Regulation Z prescribes uniform methods for computing the cost of credit, for disclosing credit terms, and for resolving errors on certain types of credit accounts. More specifically, Section 205.11 of Regulation E and Section 226.13 of Regulation Z both prescribe methods of resolution of errors (of different sorts) by the payment provider. The Federal Reserve has not yet interpreted these regulations as applying to providers of personal on-line payment systems.

11. A similar set of issues arises in the context of "electronic money" and stored-value devices. See American Bar Association Task Force on Stored-Value Cards (1997, pp. 653-72) and Group of Ten (1997).

12. For more information on narrow banks, see Litan (1987).

13. The definition paraphrases the definition of a bank contained in the Bank Holding Company Act of 1956, which considers a bank an institution that "(i) accepts demand deposits or deposits that the depositor may withdraw by check or similar means for payment to third parties or others; and (ii) is engaged in the business of making commercial loans" (Bank Holding Company Act of 1956, Section 2(C)(1) codified at 12 U.S.C. 1841(c)).

14. Glass-Steagall Act, Section 21A(2) codified at 12 USCS § 378(a)(2).

15. The definition of a "deposit," given in the Federal Deposit Insurance Act, is long and involves funds held by a bank. Therefore, at least in part, it involves the judgment that the institution concerned is a bank.

16. Blinder (1995) and Freedman (2000) reach similar conclusions on the likely monetary impact of stored-value cards and other forms of "electronic cash."

17. A similar point is made by Blinder (1995).

18. Traveler's checks are an example of a privately issued nonbank liability that is included in the M1 monetary aggregate.

19. This deterioration is documented in Friedman and Kuttner (1992, 1996). Beginning in 1993, the Federal Reserve in its semiannual report to Congress reported only "ranges" for broad money and debt growth; even these were dropped in July 2000.

20. The canonical statement of this process can be found in Mishkin (1997, pp. 436-47).

21. Although most recent macroeconomic models have emphasized interest rates over monetary aggregates, the quantity of money remains a key element in some theories of monetary transmission, such as those of Fuerst (1992) and Stein (1998).

22. See, for example, Friedman (1999), Goodhart (2000), and King (1999).

23. See Woodford (2000). Friedman's (2000) rejoinder to Woodford questions this conclusion on practical grounds, and argues that such developments could still decouple monetary policy from economic activity at the margin.

References

American Bar Association Task Force on Stored-Value Cards. 1997. "A Commercial Lawyer's Take on the Electronic Purse: An Analysis of Commercial Law Issues Associated with Stored-Value Cards and Electronic Money." Business Lawyer 52, no. 2 (February): 653-727.

Ballve, Marcelo. 2001. "Cyber Currencies Spawn 21st Century Gold Rush." Salon. <http://www.salon.com/tech/wire/2001/06/17/gold/index.html> (June 17, 2001).

Bank for International Settlements. 2001. "Statistics on Payment Systems in the Group of Ten Countries." February.

Blinder, Alan S. 1995. "Statement before the Subcommittee on Domestic and International Monetary Policy of the Committee on Banking and Financial Services, U.S. House of Representatives." October 11.

Board of Governors of the Federal Reserve System. 2001. Federal Reserve Bulletin, January.

Chakravorti, Sujit, and Alpa Shah. 2001. "A Study of the Interrelated Bilateral Transactions in Credit Card Networks." Federal Reserve Bank of Chicago Emerging Payments Occasional Paper no. EPS-2001-2.

Diamond, Douglas, and Raghuram Rajan. 2001a. "Liquidity Risk, Liquidity Creation, and Financial Fragility: A Theory of Banking." Journal of Political Economy 109, no. 2: 287-327.

———. 2001b. "Banks and Liquidity." American Economic Review 91, no. 2 (May): 422-5.

Freedman, C. 2000. "Monetary Policy Implementation: Past, Present, and Future—Will Electronic Money Lead to the Eventual Demise of Central Banking?" International Finance 3, no. 2 (July): 211-28.

Friedman, Benjamin M. 1999. "The Future of Monetary Policy: The Central Bank as an Army with Only a Signal Corps?" International Finance 2 (November): 321-38.

———. 2000. "A Response to Goodhart, Freedman, and Woodford." International Finance 3, no. 2 (July): 261-72.

Friedman, Benjamin M., and Kenneth N. Kuttner. 1992. "Money, Income, Prices, and Interest Rates after the 1980s." American Economic Review 82, no. 3 (June): 472-92.

———. 1996. "A Price Target for U.S. Monetary Policy? Lessons from the Experience with Money Growth Targets." Brookings Papers on Economic Activity, no. 1: 77-146.

Fuerst, Timothy S. 1992. "Liquidity, Loanable Funds, and Real Activity." Journal of Monetary Economics 29, no. 1 (February): 3-24.

Goodhart, Charles. 2000. "Can Central Banking Survive the IT Revolution?" International Finance 3, no. 2 (July): 198-210.

Graham, Ellen. 1996. "Community Groups Print Local (and Legal) Currencies." Wall Street Journal, June 27, p. B1.

Group of Ten. 1997. "Electronic Money: Consumer Protection, Law Enforcement, Supervisory, and Cross-Border Issues." April.

Kashyap, Anil, Raghuram Rajan, and Jeremy Stein. Forthcoming. "Banks as Liquidity Providers: An Explanation for the Co-Existence of Lending and Deposit-Taking." Journal of Finance.

King, Mervyn. 1999. "Challenges for Monetary Policy: New and Old." Bank of England Quarterly Bulletin 39 (November): 397-415.

Litan, Robert. 1987. What Should Banks Do? Washington, D.C.: Brookings Institution.

Lorek, Laura. 2000. "Trade Ya? E-Barter Thrives." Inter@ctive Week. <http://www.zdnet.com/intweek/stories/news/0,4164,2615766,00.html> (August 14, 2000).

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

McAndrews, James J., and William Roberds. 1999. "Payment Intermediation and the Origins of Banking." Federal Reserve Bank of New York Staff Reports, no. 85.

References (Continued)

Mishkin, Frederic M. 1997. The Economics of Money, Banking, and Financial Markets. 5th ed. New York: Addison-Wesley.

Mulqueen, J. 2000. "Green Stamps Make Comeback on Web." Inter@ctive Week. <http://www.zdnet.com/intweek/stories/news/0,4164,2604852,00.html> (July 17, 2000).

Radford, R. A. 1945. "The Economic Organisation of a P.O.W. Camp." Economica 12, no. 48 (November): 189-201.

Stein, Jeremy C. 1998. "An Adverse-Selection Model of Bank Asset and Liability Management with Implications for the Transmission of Monetary Policy." RAND Journal of Economics 29 (autumn): 466-86.

Woodford, Michael. 2000. "Monetary Policy in a World without Money." International Finance 3, no. 2 (July): 229-60.

John Kambhu and Patricia C. Mosser

The Effect of Interest Rate Options Hedging on Term-Structure Dynamics

- Beginning in the 1990s, the short-term dynamics of the yield curve changed in ways that appear to be related to the growth of the interest rate options market.

- When interest rates change, options dealers buy or sell securities to adjust the hedging positions that they have taken to offset their options exposures. Since the early 1990s, these trades have been, in aggregate, large enough to affect market liquidity.

- The net result of this trading activity can be to push interest rates further in the direction they were moving. Such "feedback" effects can alter the shape of the yield curve, especially when changes in interest rates are large.

- For this reason, analysts should use caution in interpreting short-run movements in the yield curve as signals of future economic developments.

Research has shown that the yield curve is a reasonably good predictor of economic activity, in part because it seems to reflect expectations of future economic fundamentals such as growth and inflation.[1] Accordingly, movements in the curve in response to economic and financial shocks are typically watched closely by market participants and policymakers. However, several recent episodes of market illiquidity, most notably the crisis in the fall of 1998, have shown that disruptions to liquidity can affect the short-term dynamics of interest rates and the shape of the curve independently of fundamentals.

In this article, we study the influence of market liquidity and dynamic trading strategies on the short-run dynamics of the yield curve. Specifically, we focus on the recent behavior of intermediate-maturity interest rates for evidence of market liquidity effects arising from the hedging of interest rate options. We base our approach in large part on the hypothesis that the hedging transactions of interest rate options dealers generate systematic trading flows in the underlying fixed-income markets following a shock to interest rates.

In the interest rate options market, dealers are net writers of options, and they manage or hedge their options exposures by taking offsetting positions in fixed-income instruments such as U.S. Treasury securities and Eurodollar futures contracts. As interest rates change, the dealers must buy or sell fixed-income securities to adjust these hedge positions. Consequently, in the aggregate, these hedging transactions can potentially affect the market prices of the hedging instruments themselves, thus leading to further changes in interest rates. Although the size

John Kambhu is a vice president and Patricia C. Mosser an assistant vice president at the Federal Reserve Bank of New York.
<john.kambhu@ny.frb.org>
<patricia.mosser@ny.frb.org>

The authors thank Ken Garbade, Charles Jones, Jim Kahn, Carol Osler, Tony Rodrigues, and two anonymous referees for helpful suggestions. They also thank Stacy Bartolomeo, Sabina Golden, and Adam Kolasinski for excellent research assistance. The views expressed are those of the authors and do not necessarily reflect the position of the Federal Reserve Bank of New York or the Federal Reserve System.

of these hedging transactions in Treasuries and Eurodollar futures is usually relatively small, the transactions' systematic relationship to changes in interest rates suggests that they may still produce small but observable feedback patterns in the short-run movements of interest rates.

Our study reveals that the short-run dynamics in the intermediate maturities of the yield curve changed around 1990, with the appearance of positive feedback in weekly interest rate changes. For example, we find that after 1990, if the yield curve "bows" up at the five-year maturity, five-year interest rates are likely to rise further in subsequent weeks. The observed positive feedback is consistent with the effects of options dealers' hedging activity, and, notably, it is found only in the 1990s, after the market for interest rate options grew to a significant size. We also provide evidence indicating that the speed at which feedback effects move through the yield curve has increased in recent years. Not surprisingly, the market liquidity/positive-feedback effects are concentrated in the weeks following the largest changes in interest rates, but they are virtually nonexistent during periods of small changes in interest rates.

Our results also suggest that very short-run movements in the yield curve should be interpreted with caution, because such movements may reflect liquidity effects as well as changes in economic fundamentals. Moreover, we find that liquidity effects are likely to be larger when interest rate changes are large and thus when policymakers and market participants are most interested in monitoring yield curve movements closely. Reassuringly, the liquidity effects uncovered in this article are not long-lasting, suggesting that interpretation of yield curve movements over longer periods of time should not be affected by our findings.

Our analysis begins with a discussion of the role of liquidity risk and positive feedback in the short-run behavior of asset prices. We then consider how the hedging of interest rate options could produce liquidity effects in the medium-term segment of the yield curve, where market survey data suggest that dynamic hedging of options could have the largest impact on transaction flows and thus on market liquidity. Next, we test for evidence of liquidity effects at a weekly frequency in both the Treasury and swap (Eurodollar) yield curves. We conclude by considering our study's implications for risk management and policy.

Liquidity Risk and Positive Feedback

Market liquidity risk is the price risk associated with executing large transactions or executing transactions quickly. The risk is manifested in a sharp movement of prices against a trader when making a large purchase or sale of a security, or, in an extreme case, when a trader is unable to execute a large trade at a reasonable price.[2] Thus, market liquidity refers to the degree to which transaction flows affect asset prices in a market separately from any change in the economic fundamentals that determine asset values.

The potential for liquidity risk to affect asset prices in ways that are distinct from the role of fundamental economic and financial variables is receiving more attention from economists and policymakers. In a recent paper, Longstaff (2001) describes how liquidity-constrained traders will make investment decisions that lead to illiquidity discounts in asset prices. The increasing role of tradable securities in the intermediation of risk and the allocation of capital is also drawing more attention to the determinants and dynamics of market liquidity. Two recent Bank for International Settlements reports (1999a, b) address the importance of market liquidity in the conduct of monetary policy and highlight the role of market liquidity in the financial market disruptions in the fall of 1998.

> *The growth of the interest rate options market to significant size by the early 1990s provides us with a naturally occurring experiment to test for liquidity effects in the yield curve.*

Related literature examines the potential for positive-feedback trading to lead to sharp changes in or overshooting of asset prices. Positive-feedback traders who buy when prices rise and sell when prices fall have the potential to drive prices further in the same direction as the initial shock. Such trades occur in the presence of stop-loss risk management strategies, in the hedging of options, and as part of momentum trading strategies. Papers by Grossman (1988), DeLong et al. (1990), and Gennotte and Leland (1990) after the stock market crash of 1987 describe how positive feedback in asset prices can emerge and be self-sustaining, despite the presence of rational traders who might otherwise link market prices to the fundamentals. Although these papers have demonstrated how positive feedback can occur, until now little systematic evidence has been found.[3]

The growth of the interest rate options market to significant size by the early 1990s provides us with a naturally occurring experiment to test for liquidity effects in the yield curve. The over-the-counter interest rate options market grew from \$10 billion of outstanding contracts in 1986 to \$561 billion in 1990 and to \$3,704 billion in 1995.[4] If the options market affects market liquidity in the underlying fixed-income markets, then we may be able to find differences in the behavior of interest rates before and after 1990.

As interest rate options became widely available, they allowed market participants who had passively borne interest rate volatility risk to trade and transfer this risk to someone else. In practice, the "someone else" has proved to be the trading units of large financial intermediaries that acquire exposure to interest rate volatility by selling interest rate options to their customers and are more likely to hedge and manage volatility risk than to bear it passively. The nonlinear nature of an options exposure requires that its hedge position be adjusted as interest rates change. Thus, options dealers are exposed to market liquidity risk when executing the trades required by the these hedge adjustments. Furthermore, because dealers are generally net writers of options, they will execute similar hedge-related trades when rates change. The systematic relationship of such trades to interest rate changes presents us with an opportunity to look for signs of their impact on market liquidity in the short-run behavior of interest rates. In our analysis, we look for changes in the dynamics of interest rates around 1990 that are consistent with the predicted market impact of the dealers' options hedging in the aggregate. (More detailed information on interest rate options and how they are hedged can be found in the box.)

Interest Rate Options and Their Hedging

Over-the-Counter Interest Rate Options

Most over-the-counter interest rate options are caps and floors on the level of interest rates; the remainder are swaptions, which are options on swaps contracts. In International Swaps and Derivatives Association market surveys from the mid-1990s, caps and floors amounted to more than 80 percent of outstanding contracts, while swaptions accounted for the rest.

Caps and floors are options on future short-term interest rates, usually six-month Eurodollar rates. In an *interest rate cap (floor)* contract, the buyer receives the difference between the market interest rate and a strike rate specified in the options contract when the market rate is above (below) the strike rate, and nothing otherwise. Most cap and floor contracts are written for several years, and thus they can be thought of as a string of call (put) options on future values of Eurodollar rates over the contract period. A dealer's portfolio of caps (floors) can therefore be thought of as a book of call (put) options on all six-month *forward interest rates* along the entire yield curve (out to the maturity of the longest maturity contract, which can be as long as ten years).

A *forward interest rate* is the interest rate for a future time period as implied by the current shape of the yield curve. For example, the three-to-five-year forward rate is the two-year interest rate for the period between three to five years in the future. In particular, it is the rate agreed to today for a two-year loan commencing three years in the future. In the case of the forward Treasury rate, it can be calculated directly from the current three- and five-year Treasury rates.

In practice, dealers do not manage their options books by directly hedging every single six-month forward interest rate exposure along the entire yield curve out to ten years. For maturities beyond two or three years, they hedge longer sections of the yield curve in blocks broken at those points where the markets in the underlying securities are most liquid. For example, an options dealer might hedge all of the six-month forward rates of between three and five years in terms of a single exposure to the three-to-five-year forward interest rate. Similarly, all six-month forward rates of between five and ten years would be hedged in terms of a single exposure to the five-to-ten-year forward interest rate.

Dynamic Hedging of Options

Generally speaking, an option can be hedged by taking an offsetting position in the underlying asset, and the required size of this position varies with the price of the underlying asset. This variability of the hedge position results from the varying sensitivity of the option's value to the price of the underlying asset as its price changes. When the underlying asset price rises by a certain amount, a call option's value will increase by a smaller amount because of the possibility that the price of the underlying asset could still reverse direction by the time the contract matures, and even fall below the strike price, rendering the option worthless. As the underlying asset's price rises further, however, this prospect of a worthless outcome becomes less likely, and the option's value becomes more responsive to changes in the underlying asset's price.

This change in the price sensitivity of the option affects the size of the position in the underlying asset required to hedge the option. To compensate for the increase in the option's price sensitivity as the underlying asset price rises, the hedge position in the underlying asset must be made larger as well. Conversely, as the asset price falls, the hedge position should shrink. This adjustment of the hedge position's size as the underlying asset price changes is called *dynamic hedging*. Such adjustments involve buying the underlying security after its price has gone up and selling it after the price has fallen. This pattern of buying and selling introduces the potential for positive feedback in asset prices, as the transactions could introduce further upward (downward) pressure on prices after an initial upward (downward) shock to asset prices. See Hull (1993) for additional information on the pricing and hedging of interest rate options.

Example of a Bond Market Hedge of an Interest Rate Cap

The hedge of an interest rate cap involves a combination of long and short positions in fixed-income securities. A long position is a bond purchased with borrowed funds, while a short position is established by borrowing the bond and then selling it. The long position is closed out, or extinguished, by selling the bond and returning the borrowed funds, and the short position is closed out

Implications of the Dynamic Hedging of Interest Rate Options

The 1995 central bank survey of the over-the-counter derivatives markets (Bank for International Settlements 1996)—the first detailed look at the structure of the markets—found that dealers had sold 50 percent more U.S. dollar interest rate options to customers than they had purchased. More recent data confirm that this asymmetry has persisted over time. This imbalance between end-user supply and demand is unique in the over-the-counter derivatives markets.[5] Generally speaking, options dealers do not leave themselves exposed to the interest rate risk in their net options positions; instead, they hedge this exposure by taking offsetting positions in other fixed-income securities. Indeed, the ability of dealers to trade in a broad range of fixed-income markets probably allows them to execute hedging transactions faster and at a lower cost than other market participants, making them more willing than others to absorb the market's net demand for interest rate options. Nevertheless, the dealers' need to adjust hedge positions as interest rates

Interest Rate Options and Their Hedging (Continued)

by buying the bond and returning it to the bond's lender. The long position gains value when the bond price rises, while the short position gains value when the bond price falls because the bond can be repurchased at a price lower than its initial sale price.

An interest rate cap exposure to a forward interest rate can be hedged with a combination of bonds whose maturities straddle the maturity of the forward rate. The hedge consists of a long position in a bond whose maturity equals the beginning date of the forward rate and a short position in a bond whose maturity equals the ending date of the forward rate.

In the case of a Treasury market hedge of an interest rate cap on the three-to-five-year forward rate, the hedge consists of a long position in a three-year Treasury note and a short position in a five-year note. This position in the two notes is exposed only to forward rates of between three and five years because the long and short note positions extinguish exposures to interest rates of up to three years' maturity (see the table). In particular, the long position offsets the exposure of the short position to any interest rate that affects both notes—forward rates of three years' maturity or less. Meanwhile, the hedge position is exposed to longer maturity interest rates, because although the three-year note is not exposed to forward rates beyond three years, the five-year note does have exposure to longer term rates (up to its five-year maturity). Thus, the net hedge position is a short position in the five-year note with exposure only to forward rates in years four and five.

What happens, then, when interest rates change? An increase in interest rates causes the value of the caps to rise, increasing the dealer's exposure. (The dealer has sold caps.) At the same time, an increase in interest rates causes the prices of notes to fall. Thus, because the net hedge position is a short position in the five-year note (see the table), a fall in the value of the note due to an increase in rates will result in a gain in value of the short hedge position.[a] The overall effect is a rise in the value of the hedge position that offsets the increase in the dealer's interest rate cap exposure.

Following the increase in interest rates, dealers will readjust their hedge positions. As noted earlier, at the higher interest rate levels, the interest rate cap becomes more sensitive to further rate changes than the initial hedge position does. Thus, to maintain an appropriate hedge, dealers will increase the size of their positions in the three- and five-year notes. It is this dynamic hedging behavior that can potentially affect the prices of fixed-income securities.

Net Exposure of Hedge Position to Interest Rates
Change in Present Value Due to an Increase in Interest Rates

Time (years)	1	2	3	4	5
Forward rates	$F_{0,1}$	$F_{1,2}$	$F_{2,3}$	$F_{3,4}$	$F_{4,5}$
Impact of higher forward rates on short position in the five-year note	(+)	(+)	(+)	(+)	(+)
Impact of higher forward rates on long position in the three-year note	(-)	(-)	(-)		
Net exposure to interest rates				(+)	(+)

Notes: A rise in interest rates will cause the present value of both the three- and five-year notes to fall. This fall in value, however, leads to a gain in the short position's value because the note can be repurchased and returned to the security lender at a price lower than its initial sale price. The gains and losses from exposure to forward rates of up to three years' maturity cancel each other, leaving only the short exposure in the five-year note to forward rates in years four and five.

[a] A short position is established by borrowing a security and selling it; the position is closed out by buying back the security at the prevailing market price and returning it to the lender. When the price of the security falls, the short position gains value because the bond can be purchased at a price lower than its original sale price.

change means that their exposure to interest rate volatility risk is converted to an exposure to market liquidity risk when executing the trades required by their hedge adjustments.

For U.S. dollar interest rate options, the most liquid instruments for hedging are Eurodollar futures contracts and U.S. Treasury securities and futures (see box). Previous estimates of the total volume of interest rate options dealers' hedging activity in the markets for these securities suggest that the hedging is likely to have the largest impact on market liquidity in the medium-term segment of the yield curve (Kambhu 1998). Although hedging activity is largest at the shortest maturities, the volume of hedging *relative to turnover volume* in Eurodollar futures is largest at the intermediate-maturity segment of the yield curve (Table 1). If dealers hedge with Eurodollar futures contracts, a 25-basis-point rise in forward rates all along the yield curve would lead to hedging transactions amounting to twice the average daily turnover volume of futures contracts at maturities of between three and five years. The same 25-basis-point rise in forward rates could generate hedging transactions of about five times the daily turnover volume at maturities of five years and beyond.[6]

The potentially large hedging impact at three-to-ten-year maturities in the Eurodollar futures market suggests that dealers will also be hedging in other liquid markets, most likely in the U.S. Treasury market. If options dealers were hedging in these markets, a 25-basis-point increase in forward rates would cause a hedge adjustment amounting to 7 percent of the daily turnover in the Treasury futures and cash markets of five- and ten-year-maturity securities (Kambhu 1998). A 75-basis-point increase in rates would cause hedge adjustments amounting to 21 percent of Treasury futures and cash turnover. Although these are not extraordinarily large shares of the Treasury market, they may be large enough to produce observable patterns in the behavior of the intermediate-maturity segment of the Treasury yield curve.[7]

Table 1

Estimated Volume of Hedging Activity Relative to Daily Turnover in the Eurodollar Futures Market

	Volume as a Percentage of Turnover	
Maturity	For a 25-Basis-Point Rate Change	For a 75-Basis-Point Rate Change
Zero to one year	5	24
One to three years	29	93
Three to five years	201	591
Five to ten years	513	1510

Source: Kambhu (1998).

At shorter maturities (less than three years), dealers' hedging of interest rate options is less likely to affect market liquidity. The shorter maturity fixed-income markets have substantially larger turnover volume and greater liquidity, and thus dealers' hedging activity can be easily accommodated. Therefore, we would not expect to find feedback effects at the short end of the yield curve.

Implications for Intermediate-Maturity Interest Rates

Our review of hedging activity thus far suggests that we look for market liquidity effects in the three-to-ten-year segment of the term structure. In particular, we will look at how the five-year spot rate reacts to past changes in the three-to-five-year forward rate and the five-to-ten-year forward rate. These maturities were chosen because data on three-, five-, and ten-year Treasury rates are available for a relatively long time period and because the liquidity of these securities makes them attractive instruments for options hedging. We focus on the five-year spot rate and forward rates on either side of five years because the hedging of exposures to either of these forward rates will require trading in a five-year security (see box).

We formulate a testable hypothesis using the hedging of interest rate caps, because caps form the bulk of the over-the-counter interest rate options market. As described in the box, the hedging of a three-to-five-year forward interest rate cap involves taking a short position in the five-year note (as well as a long position in the three-year note). This hedge position is adjusted dynamically as interest rates change because the option's value increases at an escalating rate as forward rates rise. For example, a rise in the forward rate will require a larger short position in the five-year note. If many dealers attempt to sell short the five-year note at the same time, their actions could exert downward pressure on the price of the five-year Treasury note, translating into an increase in its yield.

For an interest rate cap on the five-to-ten-year forward rate, the hedge involves a long position in the five-year note (and a short position in the ten-year note) that must also be adjusted dynamically. Consequently, a rise in the five-to-ten-year forward rate will require a larger long position in the five-year note, which in turn will lead to additional purchases of the five-year note. A large quantity of such purchases could place upward pressure on five-year note prices and thus exert downward pressure on five-year rates. Because of sizable transaction costs, we assume that such hedge adjustments are not instantaneous, but instead occur over a number of days. Below, we look for such effects at a weekly frequency.

The combination of these hedging transactions gives us the following testable hypothesis: a rise in the three-to-five-year forward rate will lead to a future rise in the five-year spot rate, and a rise in the five-to-ten-year forward rate will be followed by a fall in the five-year spot rate. We test for the presence of this relationship using the following equation:

$$(1)\quad \Delta r5 = c + \alpha Z(-1) + \beta_1 \Delta F_{3,5}(-1,-t) + \beta_2 \Delta F_{5,10}(-1,-t) + \beta_3 \Delta r5(-1,-t) + \varepsilon ,$$

where

$r5$ is the five-year interest rate;

$F_{3,5}$ is the three-to-five-year forward rate;

$F_{5,10}$ is the five-to-ten-year forward rate;

$\Delta r5 = r5 - r5(-1)$ is the one-week change in the five-year interest rate;

$\Delta F_{g,h}(-1,-t) = F_{g,h}(-1) - F_{g,h}(-t)$ is the t-1-week change in the forward rate, lagged one week, where the subscripts denote the interval of the forward rate;

$\Delta r5(-1,-t) = r5(-1) - r5(-t)$ is the t-1-week change in the five-year spot rate, lagged one week; and

$Z(-1) = r5(-1) - c - a\,F_{3,5}(-1) - bF_{5,10}(-1)$ is an error-correction term from the cointegration relationship in the levels of the five-year and forward rates, lagged one week.[8]

If hedging by options dealers has an impact on the five-year interest rate, as hypothesized above, the coefficients on the changes in forward rates will have the signs $\beta_1 > 0$ and $\beta_2 < 0$.[9]

Were the 1990s Different from Earlier Periods?

If the hedging of interest rate options affects intermediate-maturity interest rates, then the behavior of these rates in the 1990s should be different from their behavior in earlier decades. To look for this change, we begin with an analysis of Treasury interest rates. We then examine the relationship between spot rates and lagged forward rates during the 1990s using forward Eurodollar rates from the interest rate swap yield curve.

Despite the fact that the Eurodollar rate is the benchmark rate in the caps market and dealers may first look to the Eurodollar futures market to hedge their options exposures, two factors drive us to use Treasury rates initially. First, a sufficiently long sample period for the years before the 1990s is available only with Treasury interest rates, and second, arbitrage opportunities between Eurodollar rates and Treasury rates make the two interest rates very highly correlated. For the analysis using Treasury rates, we end the sample period at 1999. We end it there because new issuances of three-year Treasury notes were discontinued in 1998, and because the relationship between long-term Treasury rates and other market rates appears to have changed in response to the Treasury buy-back program, which began in early 2000. (Our analysis using swaps rates, however, employs a data set that includes 2000.)

Results Using Forward Rates from the Treasury Curve

To determine whether a change in the dynamics of interest rates occurred around 1990, we estimate regressions of equation 1 in sets of seven-year sample periods, rolling forward in one-year increments from 1965 to 1999. The data consist of weekly interest rates, with observations on the Wednesday of each week.[10] Each regression estimates the relationship between changes in the spot five-year interest rate and past changes in the three-to-five-year forward rate and the five-to-ten-year forward rate. We use two versions of these regressions, one with two-week changes in forward rates and the other with five-week changes in forward rates. (Additional details on data and estimations can be found in Appendix A.)

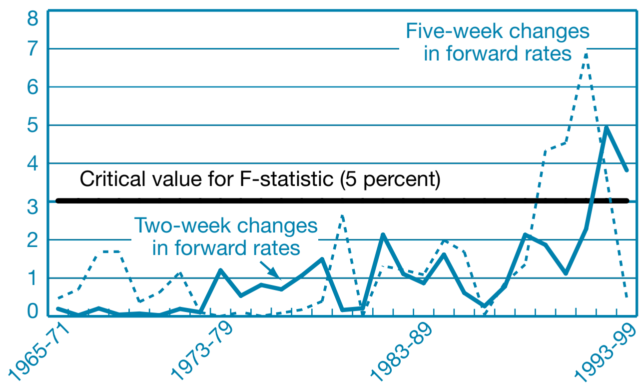
If the hedging of interest rate options affects intermediate-maturity interest rates, then the behavior of these rates in the 1990s should be different from their behavior in earlier decades.

The estimation results are summarized in Charts 1 and 2. Chart 1 depicts the statistical significance of the relationship between the spot and lagged forward rates, while Chart 2 depicts the direction and size of the relationship. The charts show a distinct difference in the relationship before and after 1990.

Chart 1 measures the strength of the statistical relationship between the spot and lagged forward rates over the 1965-99 period. It plots two lines: one for the effects of two-week changes in forward rates and the other for the effects of five-week changes in forward rates. The first point on each line is from the first rolling regression (1965-71) and the last point is for the final regression (1993-99). The test statistic shown is an F-statistic for the joint distribution of the coefficients for the two forward rates, β_1 and β_2. The F-statistic is nonstandard and is above the critical value only if both of the forward rates are statistically significant predictors of the change in the five-year rate. (For more on the test, see Appendix A.)

Chart 1

F-Statistics for the Joint Distribution of the Coefficients β_1 and β_2 from 1965 to 1999



Notes: The dependent variable is the Treasury rate; the explanatory variables are forward rates from the Treasury curve. F-statistics are from twenty-nine rolling regressions, each with seven-year sample periods. The F-statistic is for the test $H0$: $\beta_1 = 0$ or $\beta_2 = 0$; $H1$: $\beta_1 \neq 0$ and $\beta_2 \neq 0$.

Chart 1 shows that a statistically significant relationship between the spot rates and both forward rates appears only during the 1990s, and that the equations with five-week changes in forward rates are significant in the early 1990s while those with two-week changes in forward rates are significant in the latter part of the 1990s. These results suggest that changes in forward rates affect subsequent spot rates only during the 1990s, and that the impact occurs faster in the latter part of the decade. Later, we explore further the speed and duration of the forward rates' influence on the spot rate.

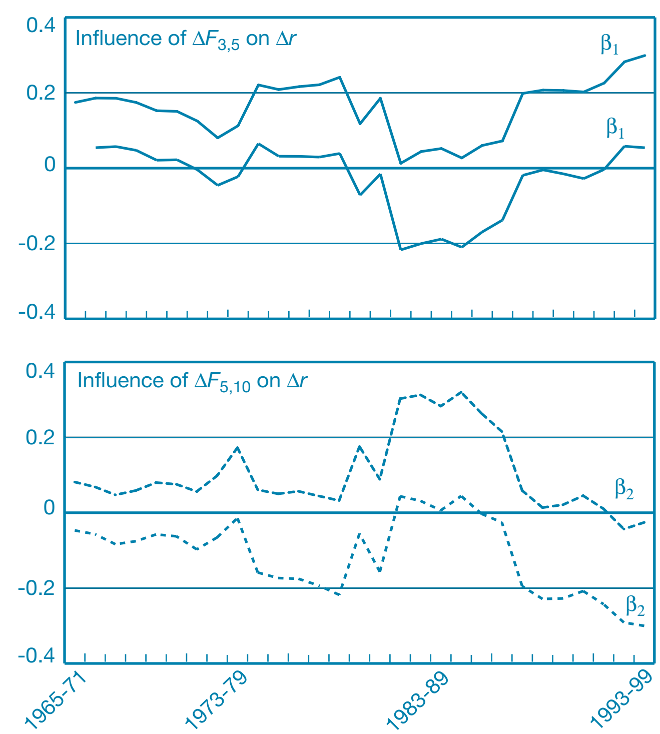
While Chart 1 summarizes only the statistical significance of the relationship between the spot and forward rates, Chart 2 shows the size and direction of the relationship. That chart provides confidence bands for the coefficients in the regressions with the two-week change in forward rates. Both coefficients are significant simultaneously and have signs consistent with hedging-related liquidity effects only after 1992 (a positive coefficient for $\Delta F_{3,5}$ and a negative coefficient for $\Delta F_{5,10}$).[11] For the periods before 1992, the lagged three-to-five-year forward rate occasionally has a significant positive effect on the five-year spot rate, but the change in the five-to-ten-year forward rate is rarely significant, and then only with a positive coefficient rather than the postulated negative coefficient.[12] Regressions using five-week changes in forward rates (not shown) produced similar results, with the notable exception being that both forward rates were significant only during the early 1990s. (The statistics from a set of representative regressions are provided in Appendix B.)

Although the lagged forward rates exert a statistically significant effect on the spot rate, they explain only a small part of the variation in the spot rate. This is evident from the small R^2 of the regression equations in Appendix B. The low explanatory power of our regressions is reflected in the fact that our regression results cannot be used to create profitable trading strategies in the presence of transaction costs. Our regression results suggest that the five-year Treasury yield can be predicted using information on past changes in forward rates. Thus, one might assume that there are unexploited profitable trading opportunities in the five-year-note market. In light of the depth and liquidity of the U.S. Treasury market, however, the existence of such unexploited profit opportunities would seem unlikely. Indeed, trading strategies based on our regression results were not consistently profitable.[13] Other implications of this low explanatory power are examined later.

Chart 2

Confidence Bands for the Coefficients β_1 and β_2 from 1965 to 1999

For Two-Week Changes in Forward Rates



Notes: The dependent variable is the Treasury rate; the explanatory variables are forward rates from the Treasury curve. Depicted are 95 percent confidence bands (one-tailed) of the coefficients from twenty-nine rolling regressions, each with seven-year sample periods.

To explore further whether these results are related to market liquidity effects, we estimated regressions similar to those above using shorter maturity interest rates. We reasoned that the much larger turnover volume and greater liquidity of the shorter maturity fixed-income markets should easily accommodate options hedging activity with little effect on shorter maturity interest rates. Indeed, regressions of changes in the one-year spot rate on lagged changes in the six-month-to-one-year forward rate and the one-to-two-year forward rate produced no evidence of positive feedback.[14]

Results with Forward Rates from the Interest Rate Swap Yield Curve

Because the predominant benchmark rate in the caps market is the Eurodollar rate, we would expect positive-feedback effects, if any, to appear first in the Eurodollar market. Thus, we repeat the analysis above using forward rates derived from the interest rate swap yield curve. The swaps data consist of one-week changes in rates from Wednesday to Wednesday, where the rates are the fixed rate in fixed-for-floating Eurodollar interest rate swaps. Because reliable data on rates of long-dated swaps are available only from the late 1980s, our sample begins in 1989. As before, our analysis consists of a series of regressions, each with a seven-year sample period rolling forward in one-year increments. (More information on our data and sources can be found in Appendix A.)

The results found using forward Eurodollar rates are the same as those found using the Treasury forward interest rates. The confidence bands for the estimated coefficients β_1 and β_2 for two-week changes in forward rates are presented in Chart 3. The top panel shows the effect on five-year Treasury rates and the bottom panel shows the effect on five-year swap rates, where the explanatory variables in both cases are forward Eurodollar rates from the swap curve. As before, the three-to-five-year forward rate has a positive coefficient and the five-to-ten-year forward rate has a negative coefficient. The two-week changes in forward rates are always statistically significant in the latter part of the decade. Similar results are obtained with the five-week changes in forward rates (not shown), except that statistically significant coefficients are found only in the earlier part of the decade.

Like the influence of Treasury forward rates, the influence of forward Eurodollar rates on the spot rate is also consistent with the predicted effect of the dynamic hedging of options.

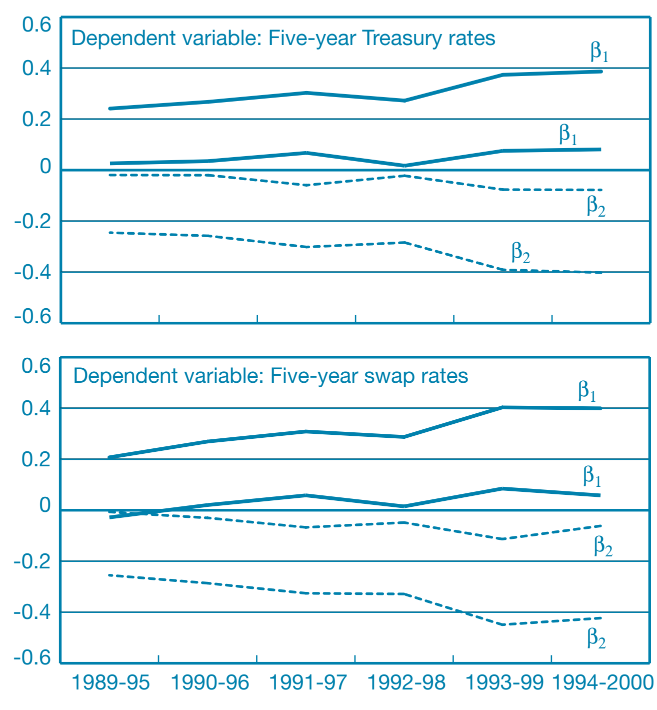
The regression results in Appendix B provide additional details on our finding that changes in forward rates affect spot rates faster toward the end of the decade. Moreover, the impact also appears to have remained at least as strong. In fact, in all the representative regressions in Tables B1 and B2 of Appendix B, the estimated coefficients for the two-week changes in forward rates in the second half of the decade are consistently larger than the coefficients for the five-week changes in forward rates in the first half of the decade.

One interpretation of the faster impact of changes in the forward rates toward the end of the 1990s is that options dealers adjusted their hedge positions faster at the end of the decade than they did at the beginning. This change in behavior could be due either to lower transaction costs or a recognition that delayed hedge adjustments, requiring larger transactions, can be costly and difficult to execute, because they strain market liquidity. Although the classic options pricing models assume continuous rebalancing of hedge positions, in practice

Chart 3

Confidence Bands for the Coefficients β_1 and β_2 for Forward Rates from the Interest Rate Swap Curve from 1989 to 2000

For Two-Week Changes in Forward Rates



Notes: Depicted are 95 percent confidence bands (one-tailed) of the coefficients from six rolling regressions, each with seven-year sample periods. β_1 is the influence of $\Delta F_{3,5}$ on Δr; β_2 is the influence of $\Delta F_{5,10}$ on Δr.

options dealers face a trade-off in the timing of their hedge adjustments. Faster or more frequent hedge adjustments produce hedge positions that more effectively match an option's exposure to price risk, but they do so at the cost of higher cumulative transaction costs over the life of the option (see, for example, Toft [1996]). Thus, a change in the trade-off from lower transaction costs would lead to faster or more frequent hedge adjustments.

Only the Big Changes Matter

Facing transaction costs, dealers might choose to adjust their hedge positions only for changes in forward rates above a certain threshold. If this asymmetry in behavior is present, then regressions with subsamples of small and large changes in forward rates should produce different results. The samples for these regressions are constructed by partitioning the data into subsamples of roughly equal size, with smaller changes in forward rates in one subsample and larger changes in the other.[15] The results of these regressions confirm the presence of the asymmetric behavior (Table 2). The feedback effects from forward rates to spot rates appear to be present only in large changes in forward rates. Further evidence that feedback effects are stronger when rate changes are large is presented in Table B3, where the effect is estimated for periods of large sustained changes in the five-year rate.

How Long Does It Last?

If our empirical results are due to hedging-related liquidity effects, we would expect them to be relatively short-lived. To examine the duration of the influence of forward rates on the spot rate, we estimate a series of regressions using changes in forward rates ranging from one to thirteen weeks. These regressions estimate the effect of changes in forward rates of up to thirteen weeks on the one-week change in the spot Treasury rate. (Appendix A provides further details of the estimation.) The results are summarized in Chart 4, which shows the strength of the statistical significance of the relationship between the spot and forward rates, and Chart 5, which depicts the direction and size of the relationship.

Chart 4 presents F-statistics measuring the statistical significance of the changes in forward rates for changes ranging from one to thirteen weeks. Results are shown for regressions using both the five-year Treasury rate and the five-year swaps rate as the dependent variable. In both regressions, the explanatory variables are forward rates from the swap curve. The regressions are estimated separately for the periods 1990 to 1995 (top panel)

Table 2

Regressions with Large and Small Changes in Forward Rates

	Large Changes $\lvert \Delta F_{3,5} \rvert \geq 26\ bp$ or $\lvert \Delta F_{5,10} \rvert \geq 26\ bp$	Small Changes $\lvert \Delta F_{3,5} \rvert < 26\ bp$ and $\lvert \Delta F_{5,10} \rvert < 26\ bp$
$Z(-1)$	0.012	0.003
	(P=0.102)	(P=0.607)
$\Delta F_{3,5}$ (-1,-6)	0.157*	-0.013
	(P=0.003)	(P=0.873)
$\Delta F_{5,10}$ (-1,-6)	-0.120*	-0.110
	(P=0.035)	(P=0.184)
R^2	0.046	0.011
Adjusted R^2	0.036	-0.001
Number of observations	276	298

Source: Authors' calculations.

Notes: The sample period is 1990-2000. The dependent variable is the Treasury rate; the explanatory variables are forward rates from the swap curve. An asterisk indicates that the coefficient is significant at the 5 percent level and with the predicted sign. The term Z in each regression is estimated using the sample in Table B2: 1989-2000. The estimated cointegrating equation is $Z = r5 - 0.032 - 7.698\ F_{3,5} + 7.007\ F_{5,10}$.

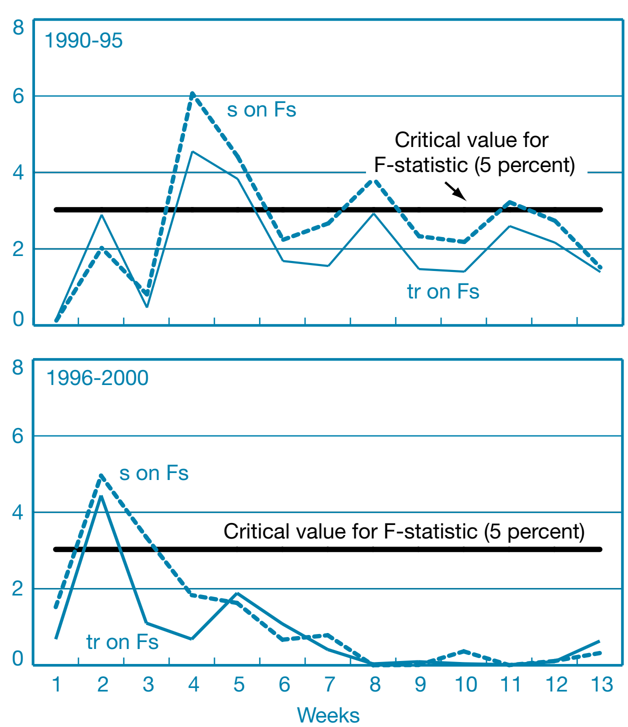
and 1996 to 2000 (bottom panel) to allow for differences between the beginning and end of the 1990s.[16] The results suggest that the statistical significance of the influence of forward rates on the spot five-year rate weakens after six weeks. In addition, a comparison of the two panels further supports our finding that the influence of lagged forward rates on the spot rate occurs more quickly in the second half of the decade; moreover, the effect appears to dissipate faster.

Chart 5 gives the confidence bands for the coefficients β_1 and β_2 for the regressions with spot Treasury rates as the dependent variable for the 1996-2000 period. The confidence bands are estimates of the size and direction of the influence of lagged changes in forward rates on the spot five-year rate. The chart shows that a two-week change in forward rates influences the Treasury spot rate in the predicted direction (a positive effect for $F_{3,5}$ and a negative effect for $F_{5,10}$). In the estimates using the full sample (top panel), the effect has weakened enough to no longer be statistically significant by about three weeks, and by nine weeks, the effect has disappeared. In a sample restricted to large rate changes (bottom panel), the feedback effects are larger and longer lasting, remaining statistically significant for as long as five to six weeks. As in the full sample, the effects disappear by about nine weeks.

The absence of a long-lasting influence of the forward rates on the five-year spot rate in these results suggests that we are observing market liquidity effects. In particular, the results are consistent with prices returning to prevailing levels after the market absorbs dealers' hedging transactions.

Chart 4

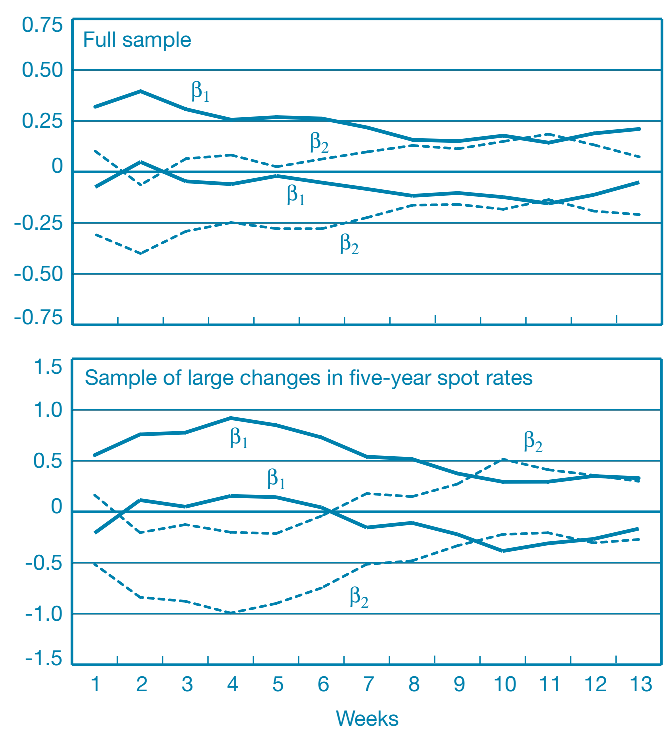
F-Statistics for the Joint Distribution of the Coefficients β_1 and β_2 for Changes in Forward Rates from One to Thirteen Weeks



Notes: The explanatory variables are forward rates from the swap curve: s on Fs is the regression of the spot swaps rate on forward rates from the swap curve; tr on Fs is the regression of the spot Treasury rate on forward rates from the swap curve. F-statistics are from thirteen regressions, each with a change in forward rates over the interval indicated on the horizontal axis—from one to thirteen weeks. The F-statistic is for the test $H0$: $\beta_1 = 0$ or $\beta_2 = 0$; $H1$: $\beta_1 \neq 0$ and $\beta_2 \neq 0$.

Chart 5

Confidence Bands for the Coefficients β_1 and β_2 for Changes in Forward Rates from One to Thirteen Weeks



Notes: The sample period is 1996-2000. β_1 is the influence of $\Delta F_{3,5}$ on Δr; β_2 is the influence of $\Delta F_{5,10}$ on Δr. The dependent variable is the spot Treasury rate; the explanatory variables are forward rates from the swap curve. Depicted are 95 percent confidence bands (one-tailed) from thirteen regressions, each with a change in forward rates over the interval indicated on the horizontal axis—from one to thirteen weeks. In the bottom panel, the regression sample is restricted to those periods in which the subsequent two-month change in the five-year Treasury rate lies within the largest 25 percent of rate changes.

Do Liquidity Effects Spill Over from the Swap Curve to the Treasury Curve?

Having analyzed the length of the relationship between lagged forward rates and the spot interest rate, we can explore the direction of the influence on the relationship between the Treasury and swap yield curves. In the regressions above, we obtain similar statistical results regardless of whether we regress five-year Treasury rates on forward Treasury rates or forward swaps rates. Similarly, forward swaps rates predict similar liquidity effects on the five-year Treasury rate and the five-year swaps rate. However, if we use forward Treasury rates to predict swaps rates, the statistical results are not as strong. Forward Treasury rates influence the five-year swaps rate only in the early part of the 1990s. In the latter half of the decade, forward Treasury rates are not statistically significant predictors of the five-year swaps rate. Thus, in the latter part of the decade, it appears that the direction of "causality" for positive-feedback effects is from the swaps market to the Treasury market.

One interpretation of this result is that the growing liquidity of the swaps market and the wider use of more refined pricing models have shifted the focus of hedging decisions toward the swap curve. At the beginning of the 1990s, more dealers may have been using earlier generations of pricing and hedging models that did not differentiate strongly between swaps rates and Treasury rates at the longer maturities. Toward the end of the decade, however, more refined pricing models based on the swap yield curve became more widely used, causing lagged Treasury forward rates to lose their explanatory power.[17]

Implications for Risk Management and Policy

At first glance, the feedback effects highlighted above appear to be second-order. The small explanatory power of the regressions indicates that the postulated market liquidity effect of the hedging of interest rate options explains only a small portion of typical fluctuations in the yield curve. In addition, the magnitude of the feedback effect is normally not large. For example, during 1996-2000, only 20 to 25 percent of the change in lagged forward rates was transmitted to the five-year spot rate (Appendix B, Tables B1 and B2). During this period, the average weekly change (absolute value) in the five-year rate was about 9.5 basis points, only 2 basis points of which could be attributed to the combined changes in the two forward rates.

If our interpretation is correct, the small impact on the yield curve is reassuring, as it implies that options dealers' liquidity risk arising from their need to adjust hedge positions should be manageable under normal circumstances. Nevertheless, the market disruptions in the fall of 1998 were a reminder that normally manageable liquidity risk can turn large in ways that surprise even experienced market participants.[18] Many institutions that relied on market liquidity to execute dynamic risk management strategies found themselves exposed to far higher risks than they had anticipated. This heightened sense of risk, in turn, caused many participants to withdraw from markets, further impairing liquidity. Although the dynamic hedging strategies examined in this article were not particularly stressed during the fall of 1998 (because the benchmark interest rate environment remained comparatively benign), the suddenness of the 1998 liquidity crisis is a warning about any market's vulnerability to dynamic risk management strategies.

The conclusion that liquidity risk is manageable under normal circumstances thus leaves room for questions about more extreme circumstances. In our empirical results, the feedback effects are stronger when interest rate changes are large (Table 2 and Chart 5). As a result, estimates based on the full sample may understate potential distortions to the yield curve during periods of large changes in interest rates. Indeed, during periods of large sustained changes in rates that continue for several weeks, the change in forward rates accounts for a relatively large proportion of the change in the five-year spot rate. For instance, if we examine periods during 1996-2000, when the five-year rate changed by more than 68 basis points over a two-month period (the largest 10 percent of such changes), more than 70 percent of the change in lagged forward rates was transmitted to the five-year spot rate (Table B3). Furthermore, changes in lagged forwards account for nearly half of the variation in the five-year spot rate during such periods. Finally, the feedback effect is also more persistent during periods of large rate changes, as we see in Chart 5.

The times when market participants and policymakers are most interested in extracting from the yield curve a signal about economic fundamentals are precisely the times when changes in the curve may be distorted by liquidity effects.

Our finding of larger feedback effects during episodes of large changes in interest rates suggests that dealers' hedging demands might run up against more severe liquidity constraints if the volatility of rates were to rise sharply. Overall, the interest rate environment of the 1990s was relatively benign by historical standards. In a more volatile environment—such as the one experienced in the late 1970s and early 1980s—dynamic hedging might introduce more disruptive positive-feedback effects if reduced market liquidity and dealers' hedging demands interacted to amplify market shocks.[19]

Furthermore, the potential for positive-feedback effects has implications for how short-run yield curve movements are interpreted by market participants and policymakers. In recent years, these movements have been followed closely for several reasons. For example, the yield curve has been shown by studies to be a relatively good predictor of economic activity, in part because it appears to reflect expectations of future economic fundamentals (see, for example, Estrella and Mishkin [1998], Estrella and Hardouvelis [1991], and Stock and Watson [1989]). In addition, the curve reflects one component of the monetary policy transmission mechanism, from short-term to long-term interest rates.

If yield curve movements over short time periods are influenced by liquidity effects as well as by expectations of economic fundamentals and policy, these movements may have to be interpreted more carefully. Yield curve changes tend to be monitored most closely when large economic and financial shocks occur or when significant policy changes are made. However, our empirical results suggest that liquidity effects in the yield curve are largest when shocks to interest rates are large. Thus, the times when market participants and policymakers are most interested in extracting from the yield curve a signal about economic fundamentals are precisely the times when changes in the curve may be distorted by liquidity effects.

Conclusion

In this article, we have examined the influence of market liquidity and dynamic trading strategies on the short-run behavior of the yield curve. Motivating our analysis was the hypothesis that dynamic hedging by sellers of interest rate options could generate transaction flows that affect market liquidity and thus produce systematic patterns in interest rate movements. The growth of the over-the-counter interest rate options market to significant size in the late 1980s allowed us to identify potential changes in interest rate dynamics that followed the development of this market. Indeed, we found a distinct difference in the dynamics of the term structure before and after 1990 that is consistent with the predicted impact of dynamic hedging of interest rate options.

Previous research on the structure of the interest rate options market has found that the largest impact of dealers' dynamic hedging on trading volume in the underlying fixed-income markets likely occurs along the intermediate-maturity section of the yield curve. For this segment of the curve, the hedging of options exposures to the three-to-five-year forward rate and the five-to-ten-year forward rate could have an effect on the five-year spot interest rate. Beginning around 1990, we find that the five-year spot rate does tend to behave as predicted following changes in the forward rates. In contrast, the relationship between the spot rate and the forward rates does not appear in the data before the 1990s.

We interpret the observed behavior of five-year interest rates as the product of short-term liquidity effects. This conclusion is based on several findings. First, the predicted relationship between forward rates and spot rates does not persist beyond a few weeks, nor can it be profitably exploited in a systematic way. Both results suggest that short-term liquidity forces rather than economic fundamentals are likely to be driving the results. In addition, and in contrast to the behavior of medium-maturity rates, shorter maturity interest rates show no evidence of such feedback effects. The ample liquidity of the markets for short-term interest rate products, where market turnover is large relative to hedging demands, makes them an unlikely site for any evidence of positive-feedback effects. Finally, forward rates predict spot rates in the medium-term segment of the yield curve only in the weeks when rate changes are relatively large. This finding is also consistent with liquidity effects, since large interest rate changes cause large adjustments to options hedges, which in turn induce trading flows that will be large relative to normal market turnover.

Although we find evidence of market liquidity effects consistent with dynamic hedging at the medium-term segment of the yield curve, the relationship accounts for only a small part of the variation in rates. The relatively small impact on the yield curve suggests that the U.S. dollar fixed-income markets are liquid and deep enough to absorb dealers' hedging transactions under normal market conditions, and that the liquidity risk arising from their need to adjust hedge positions dynamically should be manageable. However, during periods when interest rates are changing rapidly or periods of market stress when interest rate volatility jumps, liquidity effects could be significantly larger. It is exactly during such times that short-term yield curve movements may be most affected by hedge-related trading and may move in ways that are unrelated to economic fundamentals.

Appendix A: Data and Estimation Details

Data

Our analysis uses weekly changes in interest rates; we do not use daily data because transaction costs can make it uneconomical to adjust options hedges completely on a daily basis. Our choice of interest rate data is based on two criteria. First, we wish to use market rates that reflect as closely as possible the actual transaction prices at which options dealers are trading. Second, to evaluate whether interest rate dynamics have changed over time, we choose data that had a relatively long history.

For Treasury securities, these two criteria lead us to use constant-maturity Treasury rates. These rates are nearly identical to on-the-run Treasury yields (with an adjustment to maintain a fixed maturity), and they are available on a daily basis going back more than thirty years. As a check, we also performed our analysis using on-the-run interest rate data from dealer quotes reported by Bloomberg. The results were the same as those using the constant-maturity data. We elected not to use estimated zero-coupon yields, such as those in McCulloch and Kwon (1993), because such yields are based on imputed prices, not transaction prices, and because they are calculated from less liquid, off-the-run Treasury securities, which are unlikely hedging vehicles for options dealers.

For forward Eurodollar rates, we use forward rates derived from the interest rate swap yield curve. Swaps rates are the fixed rate in fixed-for-floating interest rate swaps, where the floating rate is indexed to a short-term Eurodollar interest rate (often a three-month rate). This index and the wide use of the swaps market for trading and hedging make the swap yield curve a reasonable source for forward Eurodollar rates. The rates are Reuters quotes, obtained from DRI-WEFA.

To check for day-of-week and overlapping-day effects, we also estimated equation 1 using Tuesday to (prior) Wednesday changes in rates. This alternative model specification had no effect on our results.

Estimation

Consistent with previous research, we find that the *levels* of interest rates along the yield curve are cointegrated. In other words, interest rates are generally nonstationary integrated time series, but there exists a linear combination of these rates that is stationary. As noted by Engle and Granger (1987), time-series regressions involving relationships between the changes in cointegrated variables should include a lagged cointegration term in order to control for correlation between the contemporaneous levels of the regression variables that would otherwise interfere with consistent estimation of the equation coefficients. Using standard regression techniques, we estimated a cointegrating relationship between the five-year Treasury rate and the three-to-five-year and five-to-ten-year forward rates. This relationship produced the error-correction term Z in equation 1. The cointegrating relationship was estimated with a constant term, but without a time trend, while controlling for thirteen lags of changes in rates.

In equation 1, the lagged changes in the five-year spot rate were never statistically significant in any estimation of the equation. This result suggests that the lagged changes in the five-year rate affect changes in the current five-year rate only through the lagged changes in the forward rates. (By definition, changes in the three-to-five-year and the five-to-ten-year forward rates are affected by changes in the five-year spot rate in the same observation period.) Consequently, all of our charts and tables report regression results without the lagged change in the five-year rate.

Rolling Regressions

The first set of estimates, in Charts 1 and 2, consists of twenty-nine regressions, each over seven-year sample periods rolling forward in one-year increments from 1965 to 1999. We use two versions of equation 1, corresponding to two- and five-week changes in forward rates:

$$\text{(A1)} \quad \Delta r5 = c + \alpha Z(-1) + \beta_1 \Delta F_{3,5}(-1,-3) + \beta_2 \Delta F_{5,10}(-1,-3) + \varepsilon \text{ and}$$

$$\text{(A2)} \quad \Delta r5 = c + \alpha Z(-1) + \beta_1 \Delta F_{3,5}(-1,-6) + \beta_2 \Delta F_{5,10}(-1,-6) + \varepsilon .$$

The estimation is least squares and the residuals terms are well behaved. The estimation was performed both with and without Newey-West heteroskedasticity and autocorrelation (HAC) consistent covariance estimates, with similar results produced in both cases. Charts 1-3 show the results without the Newey-West covariance estimates. The error-correction term Z in each regression is estimated using a sample that begins and ends three years before and after the regression sample (except at the end points of the full sample).

The regressions with forward rates from the swap curve (Charts 3-5 and Table 2) have similar structures. In these regressions, the error-correction term is estimated using the full sample of swaps data (1989-2000).

F-Statistics

The test for statistical significance, the results of which appear in Chart 1, is a nonstandard F-test of the joint distribution of the coefficients β_1 and β_2,

$$H0\colon \beta_1 = 0 \text{ or } \beta_2 = 0$$
$$H1\colon \beta_1 \neq 0 \text{ and } \beta_2 \neq 0 .$$

In geometric terms, the test asks whether the ninety-fifth percentile confidence ellipse of the estimated coefficients (centered on the estimated values of β_1 and β_2) intersects either of the axes $\beta_1 = 0$ or $\beta_2 = 0$. If it does not, both estimated coefficients are nonzero at a 95 percent confidence level (5 percent critical value).[20]

To perform the test, rather than finding all values of β_1 and β_2 on the confidence ellipse (that is, all combinations of β_1 and β_2 for the 5 percent critical value of the F-statistic) and seeing whether these are in the interior of the parameter space, we construct F-statistics for the estimated coefficients along the axes for β_1 and β_2 and ask whether the F-statistics exceed the 5 percent critical value. If they do, the ninety-fifth percentile confidence ellipse must be in the interior of the parameter space. Specifically, we calculate F-statistics for the joint distribution of the estimated coefficients β_1 and β_2 along the axis $\beta_1 = 0$ and the axis $\beta_2 = 0$, with the distribution centered on the estimated values of β_1 and β_2. The F-statistic chosen for the test is the smallest of these F-statistics, and the null hypothesis is rejected when this minimum value exceeds the critical value. This F-statistic is presented in Chart 1.

The Duration of the Relationship between the Spot Rate and Lagged Forward Rates

To determine the duration of the influence of the forward rates on the spot rate, we estimate the equation

$$\text{(A3)} \qquad \Delta r5 = c + \alpha Z(-1) + \beta_1 \Delta F_{3,5}(-1, -t) + \beta_2 \Delta F_{5,10}(-1, -t) + \varepsilon$$

for $t = 2, 3, \ldots, 14$, giving us a set of thirteen regressions of the effect of changes in forward rates ranging from a one-week change to a thirteen-week change. As in the earlier case, $\Delta r5$ is the one-week change in the spot Treasury rate and $Z(.)$ is the error-correction term from the cointegration relationship in the level of rates, and is estimated as described above. The test for statistical significance of the relationship between the spot and forward rates uses the same test procedure described earlier. The F-statistics for the joint distribution of β_1 and β_2 are shown in Chart 4; confidence bands for the β_1 and β_2 estimates are presented in Chart 5. The test statistics in these charts were computed with Newey-West HAC consistent covariances because of heteroskedastic residuals in the shorter sample periods, particularly 1996-2000.

Appendix B: Representative Regression Results

Results Using Treasury Interest Rates

The statistics for a set of regressions illustrating the differences between the early and later parts of the 1990s are presented in Table B1. These regressions are performed for two subperiods: the six-year period from 1990 to 1995 and the four-year period from 1996 to 1999. Two regressions are estimated in each subperiod, one with two-week changes in forward rates (equation A1) and the other with five-week changes in forward rates (equation A2).

The statistics in Table B1 show that for the early 1990s, only the five-week changes in forward rates (equation A2) are statistically significant with the anticipated signs (a positive sign for changes in the three-to-five-year forward rate and a negative sign for changes in the five-to-ten-year forward rate). For the late 1990s, only the two-week changes (equation A1) are statistically significant with the anticipated signs. These findings, plus the absence of significant results for the other regressions, are consistent with Charts 1 and 2: lagged changes in forward rates affect five-year yields more quickly in the later part of the 1990s than in the earlier years of the decade. In the earlier period, only the regression with the five-week change in forward rates has significant coefficients, while in the later period, only the regression with the two-week change in forward rates has significant coefficients.

Results Using Forward Rates from the Swap Curve

Illustrative regression statistics from equations using swaps interest rates are shown in Table B2. The explanatory variables in the regressions are forward rates derived from the swap yield curve, while the dependent variable is the spot Treasury rate in

Table B1

Regression of the Spot Treasury Rate on Forward Treasury Rates

	1990-95		1996-99	
	Equation A1: Two-Week Change in Forward Rates	Equation A2: Five-Week Change in Forward Rates	Equation A1: Two-Week Change in Forward Rates	Equation A2: Five-Week Change in Forward Rates
$Z(-1)$	0.009	0.017	0.019	0.023
	(P=0.270)	(P=0.020)	(P=0.032)	(P=0.006)
$\Delta F_{3,5}$ (-1,-3)	0.081		0.197*	
	(P=0.269)		(P=0.005)	
$\Delta F_{5,10}$ (-1,-3)	-0.080		-0.195*	
	(P=0.247)		(P=0.028)	
$\Delta F_{3,5}$ (-1,-6)		0.160*		0.079
		(P=0.001)		(P=0.148)
$\Delta F_{5,10}$ (-1,-6)		-0.123*		-0.043
		(P=0.016)		(P=0.523)
R^2	0.006	0.034	0.032	0.027
Adjusted R^2	-0.003	0.025	0.017	0.022

Source: Authors' calculations.

Notes: An asterisk indicates that the coefficient of the change in forward rates is significant at the 5 percent level (two-tailed) and with the predicted sign. Equations A1 and A2 for 1990-95 are estimated using least squares and Newey-West HAC consistent covariance, while an EGARCH(2,2) correction was used in the estimation of equations A1 and A2 for 1996-99. The term Z in each regression is estimated using the sample in Table B2: 1989-2000. The estimated cointegrating equation is $Z = r5 - 0.006 - 6.566\, F_{3,5} + 5.57\, F_{5,10}$.

Table B2
Regressions Using Forward Eurodollar Rates

	1990-95		1996-2000	
	Equation A1: Two-Week Change in Forward Rates	Equation A2: Five-Week Change in Forward Rates	Equation A1: Two-Week Change in Forward Rates	Equation A2: Five-Week Change in Forward Rates
Panel A: Regression of Spot Treasury Rate on Forward Eurodollar Rates				
$Z(-1)$	0.002	0.006	0.033	0.032
	(P=0.645)	(P=0.269)	(P=0.004)	(P=0.011)
$\Delta F_{3,5}$ (-1,-3)	0.143		0.222*	
	(P=0.070)		(P=0.036)	
$\Delta F_{5,10}$ (-1,-3)	-0.130		-0.232*	
	(P=0.090)		(P=0.024)	
$\Delta F_{3,5}$ (-1,-6)		0.141*		0.124
		(P=0.007)		(P=0.157)
$\Delta F_{5,10}$ (-1,-6)		-0.104*		-0.127
		(P=0.051)		(P=0.171)
R^2	0.012	0.029	0.041	0.033
Adjusted R^2	0.003	0.019	0.029	0.022
Panel B: Regression of Spot Swaps Rate on Forward Eurodollar Rates				
$Z(-1)$	0.014	0.020	0.079	0.076
	(P=0.260)	(P=0.089)	(P=0.015)	(P=0.032)
$\Delta F_{3,5}$ (-1,-3)	0.126		0.223*	
	(P=0.156)		(P=0.027)	
$\Delta F_{5,10}$ (-1,-3)	-0.141		-0.240*	
	(P=0.116)		(P=0.024)	
$\Delta F_{3,5}$ (-1,-6)		0.136*		0.119
		(P=0.010)		(P=0.186)
$\Delta F_{5,10}$ (-1,-6)		-0.117*		-0.122
		(P=0.036)		(P=0.203)
R^2	0.011	0.021	0.031	0.023
Adjusted R^2	0.002	0.012	0.020	0.011

Source: Authors' calculations.

Notes: An asterisk indicates that the coefficient of the change in forward rates is significant at the 5 percent level (two-tailed) and with the predicted sign. Equations are estimated using least squares with Newey-West HAC consistent covariance. The cointegration term Z in each regression is estimated using the full sample of swaps data: 1989-2000. The estimated cointegrating equations are $Z = r5 - 0.032 - 7.698\, F_{3,5} + 7.007\, F_{5,10}$ for panel A, and $Z = r5 - 0.019 - 4.575\, F_{3,5} + 3.764\, F_{5,10}$ for panel B.

panel A and the spot swaps rate in panel B. The results in these panels are similar to those arrived at using forward rates from the Treasury curve. As noted, changes in forward rates affect five-year yields more quickly toward the end of the 1990s.

RESULTS FOR PERIODS OF LARGE CHANGES IN THE FIVE-YEAR TREASURY RATE

The regressions in Table B3 estimate the relationship between the spot five-year rate and lagged forward rates in periods leading up to episodes of large changes in the five-year rate. The regression sample is restricted to periods in which the subsequent two-month change in the five-year rate lies within the largest 10 percent of rate changes. In these regressions, the changes in forward rates have a substantially stronger influence on the spot five-year rate than they do in the full-sample regressions in Tables B1 and B2. For instance, during 1996-2000, the predicted one-week change in the spot five-year rate is more than 70 percent of the two-week change in forward rates—compared with only 20 percent in the full sample. In addition, during this period, the R^2 of the regression indicates that almost half of the variability in the spot five-year rate can be attributed to the change in forward rates.

TABLE B3

Regressions for Periods of Large Changes in the Five-Year Treasury Rate

	1990-2000	1996-2000
$Z(-1)$	0.062	0.110
	(P=0.001)	(P=0.003)
$\Delta F_{3,5}$ (-1,-3)		0.720
		(P=0.057)
$\Delta F_{5,10}$ (-1,-3)		-0.956
		(P=0.008)
$\Delta F_{3,5}$ (-1,-6)	0.522	
	(P=0.000)	
$\Delta F_{5,10}$ (-1,-6)	-0.443	
	(P=0.002)	
R^2	0.26	0.41
Adjusted R^2	0.22	0.32
Average value of dependent variable (basis points)		
Predicted	8	7
Actual	13	10
Number of observations	57	25

Source: Authors' calculations.

Notes: The sample is the largest 10 percent of two-month changes in the Treasury rate in the indicated period. In the regression equation $\Delta r5 = c + \alpha Z(-1) + \beta_1 \Delta F_{3,5}(-1,-t) + \beta_2 \Delta F_{5,10}(-1,-t) + \varepsilon$, all terms are as defined in equation 1. The regression sample is defined by $|r5(8) - r5(-1)| \geq 77$ basis points for 1990-2000 and $|r5(8) - r5(-1)| \geq 68$ basis points for 1996-2000, where $r5(8) - r5(-1)$ is the leading two-month change in the spot five-year Treasury rate. All equations are estimated using least squares and Newey-West HAC consistent covariance. The explanatory variables are forward rates derived from the swap yield curve. The term Z in each regression is estimated using the sample in Table B2: 1989-2000. The estimated cointegrating equation is $Z = r5 - 0.032 - 7.698\, F_{3,5} + 7.007\, F_{5,10}$.

Endnotes

1. The yield curve depicts interest rates of different maturities at a point in time. For more on the yield curve's role as a predictor of economic activity, see, for example, Estrella and Mishkin (1998), Estrella and Hardouvelis (1991), and Stock and Watson (1989).

2. In the presence of liquidity risk, traders have a choice when making a large trade: they can either accept a disadvantageous price in return for immediate trade execution, or they can spread the trade over a series of smaller transactions and sacrifice immediacy of execution while bearing exposure to price risk until all transactions are completed.

3. An exception is Kodres (1994).

4. Figures are from International Swaps and Derivatives Association survey data.

5. See Kambhu (1998).

6. A 25-basis-point change in forward rates is slightly less than the largest daily change and represents approximately the ninetieth percentile of one-week changes in forward rates during the 1990s.

7. Other dynamic trading activity also has the potential to affect market liquidity in the fixed-income markets. For example, the hedging of mortgage-backed securities involves the dynamic adjustment of hedge positions as mortgage prepayment rates change in response to interest rate changes. Adjustments of mortgage-backed-securities hedges may have influenced the shape of the yield curve following the Federal Reserve's interest rate hikes in 1994 (Fernald, Keane, and Mosser 1994).

8. Equation 1 can be interpreted as an error-correction model for the five-year spot rate as a function of the lagged forward rates and the lagged five-year rate. Because the forward rates are functions of the three-, five-, and ten-year spot rates, the equation could be rewritten as an error-correction model of the five-year rate as a function of lagged three-, five-, and ten-year rates. To focus on the possible feedback effects of options trading, we chose to write the model in terms of forward rates. The coefficients *a*, *b*, and *c* in the cointegration term are estimated separately. See Appendix A for more details.

9. Our regression is different from that usually used for testing the expectations hypothesis of the term structure of interest rates. In the usual expectations-hypothesis framework, one-week changes in the five-year yield would be written as a function of the lagged slope of the term structure (from one week to five years) plus an (unforecastable) error term representing the one-week excess return on the five-year note (see Campbell [1995]). The lagged changes in rates should not be significant. Although our regressions are not directly comparable to the traditional expectations-hypothesis framework, we find that lagged changes in forward rates do affect changes in the five-year rate during the 1990s. Moreover, if we augment our regression by adding the lagged spread between the five-year rate and a one-week (LIBOR) rate, we still find that lagged forward rates predict the weekly change in the five-year yield during the 1990s.

10. The data are constant-maturity Treasury (CMT) yields. Forward rates are derived from the CMT rates. See Appendix A for details.

11. Although the two forward rates are correlated, their collinearity does not seem to account for the relationship depicted in these charts. The correlation between the forward rates in the same sample periods as the regressions in Charts 1 and 2 bears no relationship to the results in those charts. The periods during the 1980s in which the degree of correlation was the same as in the 1990s did not have the same regression results as in the latter period.

12. Indeed, during the 1980s, there is evidence of mild negative feedback (a negative coefficient for $\Delta F_{3,5}$ and a positive coefficient for $\Delta F_{5,10}$) from the forward rates to the five-year spot rate, although it is not statistically significant.

13. The trades consisted of either a long or short position in the five-year Treasury note, depending on the signal from the forward rates, and an offsetting duration-matched position in the two- and ten-year Treasury securities. This "butterfly" trade created a position with exposure to the five-year spot rate but without exposure to changes in the level of the yield curve (see Garbade [1996, Chapter 14] for more on butterfly trades and the yield curve). Profits were calculated net of the transaction cost of 0.75 basis point per trade. Several strategies were tested, each of which required a different level of the signal from the forward rates before a trade was undertaken. Of those tested, the best performing strategy was profitable only at the very end of the 1990s.

14. The data used were Treasury bill rates for the six-month and one-year rates and constant-maturity rates for the two-year Treasury note. We found some evidence in the late 1980s that lagged forwards had predictive power for the one-year-note rate, but the relationship proved to be the opposite of what would be expected from the liquidity effects of hedging activity.

15. The partition was achieved by increasing a cutoff value for the change in both forward rates in 1-basis-point increments until the subsample of changes in forward rates that exceeds the cutoff value was smaller than the other subsample.

16. Although the regression samples for the early and later parts of the 1990s are split at the middle of the decade (1995 and 1996), similar results are still found with other partitions. In other words, our results are robust with respect to how the decade is split into early and later periods.

17. At the very end of the decade, reductions in the supply of Treasury securities along with the Treasury buy-back program probably exacerbated this trend.

18. See Bank for International Settlements (1999b) and Fleming (2000) for measures of the degree of illiquidity in what were normally liquid markets.

19. This possibility, however, does not imply that restrictions on dynamic hedging or option-like products are warranted. Indeed, restrictions would be undesirable for two reasons. First, dynamic hedging is not disruptive under normal market conditions. Second, restrictions on financial products whose risks are managed dynamically would limit the use of financial innovations that provide benefits to a wide range of economic agents, from residential mortgage borrowers to institutional investors. A more appropriate policy and risk management response would be prudent risk-based capital levels and robust liquidity management.

20. See Dhrymes (1978, pp. 80-3) for further discussion.

References

Bank for International Settlements. 1996. "Central Bank Survey of Foreign Exchange and Derivatives Market Activity." Basel, Switzerland.

———. 1999a. "Market Liquidity: Research Findings and Selected Policy Implications." Basel, Switzerland.

_____. 1999b. "A Review of Financial Market Events in Autumn 1998." Basel, Switzerland.

Campbell, John Y. 1995. "Some Lessons from the Yield Curve." Journal of Economic Perspectives 9, no. 3 (summer): 129-52.

DeLong, J. Bradford, Andrei Schleifer, Lawrence Summers, and Robert Waldman. 1990. "Positive Feedback Investment Strategies and Destabilizing Rational Speculation." Journal of Finance 45, no. 2 (June): 379-95.

Dhrymes, Phoebus. 1978. Introductory Econometrics. New York: Springer-Verlag.

Engle, R. F., and C. W. J. Granger. 1987. "Co-Integration and Error Correction: Representation, Estimation, and Testing." Econometrica 55, no. 2 (March): 251-76.

Estrella, Arturo, and Gikas Hardouvelis. 1991. "The Term Structure as a Predictor of Real Economic Activity." Journal of Finance 46, no. 2 (June): 555-76.

Estrella, Arturo, and Frederic Mishkin. 1998. "Predicting U.S. Recessions: Financial Variables as Leading Indicators." Review of Economics and Statistics 80, no. 1 (February): 45-61.

Fernald, Julia, Frank Keane, and Patricia Mosser. 1994. "Mortgage Security Hedging and the Yield Curve." Federal Reserve Bank of New York Quarterly Review 19, no. 2 (summer-fall): 92-100.

Fleming, Michael. 2000. "The Benchmark U.S. Treasury Market: Recent Performance and Possible Alternatives." Federal Reserve Bank of New York Economic Policy Review 6, no. 1 (April): 129-45.

Garbade, Kenneth D. 1996. Fixed Income Analytics. Cambridge: MIT Press.

Gennotte, Gerard, and Hayne Leland. 1990. "Market Liquidity, Hedging, and Crashes." American Economic Review 80, no. 5: 999-1021.

Grossman, Sanford. 1988. "An Analysis of the Implications for Stock and Futures Price Volatility of Program Hedging and Dynamic Hedging Strategies." Journal of Business 61 (July): 275-98.

Hull, John. 1993. Options, Futures, and Other Derivative Securities. Englewood Cliffs, N.J.: Prentice-Hall.

Kambhu, John. 1998. "Dealers' Hedging of Interest Rate Options in the U.S. Dollar Fixed-Income Market." Federal Reserve Bank of New York Economic Policy Review 4, no. 2 (June): 35-57.

Kodres, Laura E. 1994. "The Existence and Impact of Destabilizing Positive Feedback Traders: Evidence from the S&P 500 Index Futures Market." Board of Governors of the Federal Reserve System Finance and Economics Discussion Series no. 1994-9, May.

Longstaff, Francis. 2001. "Optimal Portfolio Choice and the Valuation of Illiquid Securities." Review of Financial Studies 14, no. 2 (summer): 407-31.

McCulloch, J. H., and H. C. Kwon. 1993. "U.S. Term Structure Data, 1947-1991." Ohio State University Economics Working Paper no. 93-6, March.

Stock, James H., and Mark W. Watson. 1989. "New Indexes of Coincident and Leading Economic Indicators." In Olivier Blanchard and Stanley Fischer, eds., NBER Macroeconomics Annual: 1989, 351-94. Cambridge: MIT Press.

Toft, Klaus Bjerre. 1996. "On the Mean-Variance Tradeoff in Option Replication with Transactions Costs." Journal of Financial and Quantitative Analysis 31, no. 2 (June): 233-63.

THE FEDERAL RESERVE BANK OF NEW YORK
ECONOMIC POLICY REVIEW

Single-copy subscriptions to the *Economic Policy Review* (ISSN 0147-6580) are free. Multiple copies are available for an annual cost of $12 for each additional subscription. Checks should be made payable in U.S. dollars to the Federal Reserve Bank of New York and sent to the Public Information Department, 33 Liberty Street, New York, NY 10045-0001 (212-720-6134). Single and multiple copies for U.S. subscribers are sent via third- and fourth-class mail.

Library of Congress Card Number: 77-646559