Federal Reserve Bank of St. Louis

REVIEW

MAY/JUNE 2005 VOLUME 87, NUMBER 3

Review is published six times per year by the Research Division of the Federal Reserve Bank of St. Louis and may be accessed through our web site: **research.stlouisfed.org/publications/review**. All non-proprietary and nonconfidential data and programs for the articles written by Federal Reserve Bank of St. Louis staff and published in *Review* also are available to our readers on this web site. These data and programs are also available through Inter-university Consortium for Political and Social Research (ICPSR) via their FTP site: www.icpsr.umich.edu/pra/index.html. Or contact the ICPSR at P.O. Box 1248, Ann Arbor, MI 48106-1248; 734-647-5000; **netmail@icpsr.umich.edu**.

Single-copy subscriptions are available free of charge. Send requests to: Federal Reserve Bank of St. Louis, Public Affairs Department, P.O. Box 442, St. Louis, MO 63166-0442, or call (314) 444-8808 or 8809.

General data can be obtained through FRED (Federal Reserve Economic Data), a database providing U.S. economic and financial data and regional data for the Eighth Federal Reserve District. You may access FRED through our web site: **research.stlouisfed.org/fred**.

ISSN 0014-9187

An Introduction to Two-Rate Taxation of Land and Buildings

Jeffrey P. Cohen and Cletus C. Coughlin

When taxing real property at the local level in the United States, land and improvements to the land, such as buildings, are generally taxed at the same rate. Two-rate (or split-rate) taxation departs from this practice by taxing land at a higher rate than structures. This paper begins with an elementary discussion of taxation and the economic rationale for two-rate taxation. In theory, moving to a two-rate tax reduces the deadweight losses associated with distortionary taxation and generates additional economic activity. The paper also provides a history of two-rate taxation in the United States and a summary of studies attempting to quantify its economic effects. Discussions of the practical and political challenges of implementing two-rate taxation complete the paper.

Federal Reserve Bank of St. Louis *Review*, May/June 2005, *87*(3), pp. 359-74.

"In my opinion, the least bad tax is the property tax on the unimproved value of land, the Henry George argument of many, many years ago."

—Milton Friedman,
as quoted in Mankiw (2004),
1976 Nobel Prize laureate in economics.

"The property tax is, economically speaking, a combination of one of the worst taxes—the part that is assessed on real estate improvements… and one of the best taxes—the tax on land or site value."

—William Vickrey (1999),
1996 Nobel Prize laureate in economics.

Revenues from the taxation of real property play a key, and frequently controversial, role in the funding of elementary and secondary education as well as many other publicly provided services. Our focus is on one suggested improvement of property taxation known as "two-rate" or "split-rate" taxation. When taxing a specific parcel of real property in the United States, the same rate is usually applied to the land as well as to the improvements to the land, such as buildings. The opinions expressed by Nobel Prize winners Milton Friedman and William Vickrey are at the root of proposals to differentiate the taxing of land from the buildings on that land. Such proposals have attracted increasing attention from researchers and policymakers in Connecticut, Massachusetts, Virginia, and Pennsylvania in recent years.[1]

The two-rate proposal is a modification of the extreme case in which the only tax levied is on the value of land. Taxes on the value of buildings, as well as all other taxes, would be zero. Thus, the owners of buildings would no longer pay a property tax; only the owners of land would be taxed. Such a pure land tax approach was advocated by Henry George in a book published in 1879, *Progress and Poverty*. He argued that land should be taxed at 100 percent of its "rental value."[2] George reasoned that the land value tax

[1] Craig (2003) notes that over 700 cities worldwide use two-rate taxation of property.

[2] O'Sullivan (2003, p. 154) defines land rent as "the annual payment in exchange for the right to use the land."

Jeffrey P. Cohen is an assistant professor of economics at the Barney School of Business, University of Hartford; he acknowledges the support of the Lincoln Foundation. Cletus C. Coughlin is deputy director of research and a vice president at the Federal Reserve Bank of St. Louis. Jason Higbee provided research assistance. The views expressed are those of the authors and do not necessarily reflect official positions of the Federal Reserve System or the Lincoln Foundation.

would increase efficiency (and the wealth of society) by allowing governments to abolish taxes on improvements to land, as well as eliminate all other forms of taxation. He also cited equity reasons for the "single tax" on land.[3] Namely, increases in land value (exclusive of improvements) in the early 1900s were due primarily to an entire community's private-sector and public-sector economic activities rather than the actions of the specific land owner. Therefore, George argued, land owners should not benefit disproportionately from city growth, and the tax on land would allow for the redistribution of these unearned gains.[4]

A pure land tax, however, is not without some faults. First, it is not easy to measure the value of land net of improvements, and this would make it difficult for government to determine the amount of the land tax. This shortcoming of the pure land tax would also be present with the two-rate tax. Second, if a pure land tax were to capture all current and future rent from the landowners, the market value of the land would become zero. This would be equivalent to the government taking the land from landowners. Thus, people would have no incentive to hold land, leading to abandonment of the land, and likely resulting in governmental decisions about how the land should be used and by whom. Third, the change to a pure land tax would likely have significant redistributional effects, with large landowners likely incurring substantial adverse wealth effects. This effect, however, might be viewed by some as a virtue rather than a fault.

Compared with the pure land tax, the two-rate tax on land and buildings is a more general and, perhaps, more practical alternative. Instead of taxing land and structures at the same rate, as is the case with the conventional property tax, the two-rate tax would tax land at a higher rate than the structures on the land. This form of the tax would encourage improvements on relatively small lots of land because such improvements would be taxed at a lower rate than the land itself. The increased incentive for improving structures would lead to increased economic development. In the context of urban economic development, such a tax policy might reverse the trend of economic decay experienced by some cities. Moreover, because the tax rate on land using a two-rate tax would be less than the tax rate using a pure land tax, the potentially large changes in land prices would be mitigated somewhat. Thus, the size of adverse wealth effects would be reduced, which might help in reaching a political agreement to support a two-rate system.

In the next section, we provide an introduction to taxation. This general introduction provides the foundation for a discussion of land taxation and the two-rate tax. A history of two-rate taxation in the United States follows this discussion. Two-rate taxation has been used in Pennsylvania, most notably in Pittsburgh. Next, we review a number of studies attempting to quantify the effects of two-rate taxation. These studies include a case study of Pittsburgh as well as studies attempting to identify the potential effects of various tax-change proposals. This discussion is followed by an elaboration of the practical problems of implementing two-rate taxation. A summary of the political economy issues involved in land value taxation completes the body of our paper.

THE WELFARE EFFECTS OF TAXATION USING DEMAND AND SUPPLY CURVES[5]

Demand and supply curves can be used to illustrate how taxes affect the behavior and the economic well-being of consumers and producers. The effect of taxation is one of the topics in what economists refer to as welfare economics. Before

[3] George used the term "single tax" because he thought that a tax on unimproved land could yield sufficient revenues to finance all government spending. While his thinking was appropriate for the late nineteenth century, the revenue potential of land taxation is far less than the size of public spending today.

[4] A discussion of the ethical arguments involving land taxation (i.e., Is such a tax just?) can be found in Fischel (1998) and Bromley (1998).

[5] Readers familiar with this material might want to proceed directly to the next section on land taxation. For readers desiring an elaboration of the material in this section, see Chapters 6 through 8 in Mankiw (2004).

Figure 1

Demand and Consumer Surplus

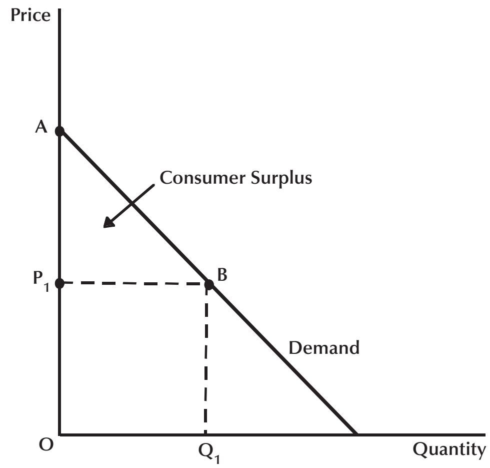
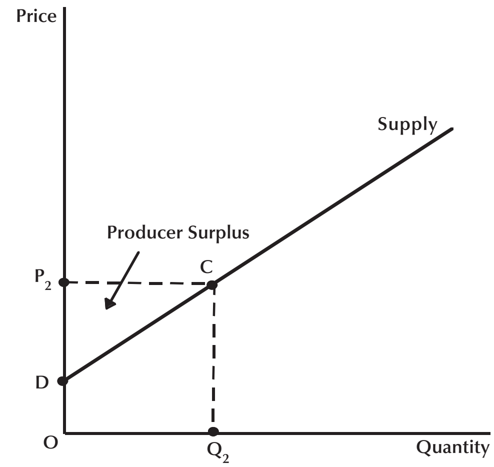


Figure 2

Supply and Producer Surplus



analyzing the effects of a specific tax, we must first introduce the concepts of consumer surplus, producer surplus, and efficiency.

Consumer Surplus

We begin by using a demand curve to measure consumer surplus. Consumer surplus is the amount that buyers are *willing* to pay for a good minus the amount they *actually* pay for it. Figure 1 shows a hypothetical demand curve. Assuming a price per unit of P_1, buyers would purchase Q_1 units of this good. Thus, the total expenditure by consumers on this good would be P_1 times Q_1 or, in terms of an area, the rectangle OP_1BQ_1.

The demand curve reveals how much consumers are willing to pay for the Q_1 units. Moving rightward on the quantity axis from the origin, the value that consumers are willing to pay is reflected by the height of the demand curve. This reflects the assumption that the first unit is valued the most by consumers, the second unit somewhat less, and so on as one moves down the demand curve. Thus, the total amount that consumers are willing to pay for the Q_1 units is equal to the area of the four-sided figure $OABQ_1$. The difference between what consumers are willing to pay and what they actually pay (i.e., consumer surplus) is represented by the triangular area P_1AB.

Producer Surplus

Turning to the supply side of this market, the concept of producer surplus can be illustrated using a supply curve. Producer surplus is the difference between *actual* compensation received by sellers for a given quantity of output (i.e., total revenue) and the minimum compensation sellers *would require* to provide that given quantity. Figure 2 shows a hypothetical supply curve. The positive slope of the supply curve reflects the assumption that the compensation to induce additional units of production increases as output increases. Assuming a price per unit of P_2, producers would supply Q_2 units of this good. Thus, the total revenue would be P_2 times Q_2 or the area of the rectangle OP_2CQ_2.

The minimum compensation necessary to induce producers to supply Q_2 is revealed by the supply curve. Moving rightward on the quantity axis from the origin, the minimum compensation is reflected by the height of the supply curve. Thus, the minimum compensation for Q_2 units is equal to the four-sided area $ODCQ_2$. The difference between the total revenue of producers and the minimum they must receive to produce (i.e.,

Figure 3

Market Equilibrium: Consumer and Producer Surplus

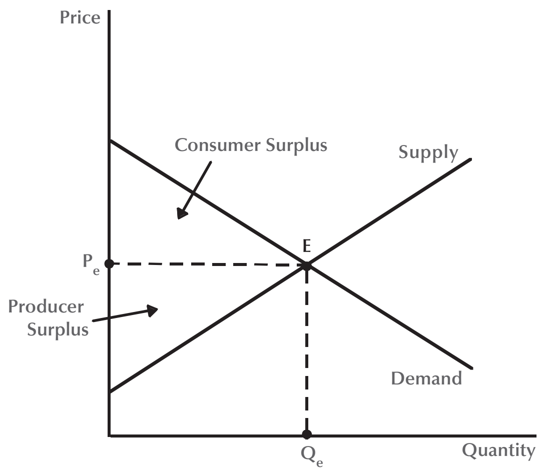
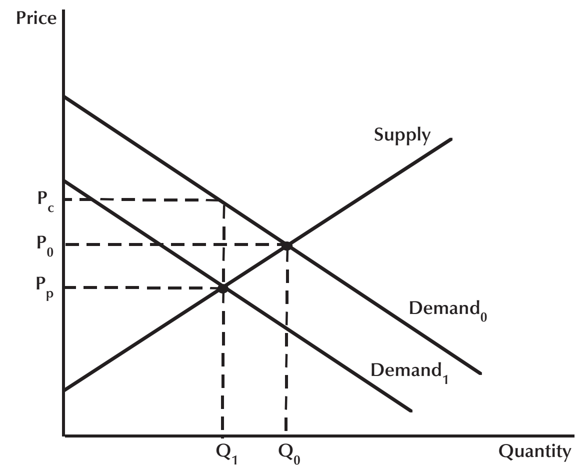


Figure 4

Effects of a Tax Levied on Consumers



producer surplus) is represented by the triangular area DP_2C.

In equilibrium, the quantity consumers are willing to buy equals the quantity producers are willing to supply. The competitive equilibrium shown in Figure 3 reflects what economists term "economic efficiency." One key implication of economic efficiency is that total economic surplus, which is the sum of consumer and producer surplus, cannot be increased by either increasing or decreasing the output of this good. Increased output would cause the additional cost incurred by sellers to exceed the additional value to buyers, while decreased output would cause the additional value to buyers to exceed the additional cost incurred by sellers.

Effects of Taxes

Now let's examine the effect of taxes on this market. Taxes impose a wedge between what consumers pay and what producers receive. Assume a tax of $1 per unit is levied on buyers of a particular good. Such a tax can be illustrated by shifting the demand curve downward by the size of the tax because consumers only care about their out-of-pocket expense in determining a desired quantity. This is shown in Figure 4 by the shift in demand from Demand_0 to Demand_1. The new equilibrium quantity is Q_1, which is less than the original equilibrium of quantity of output. At Q_1, the price per unit received by producers is P_p and the price paid by consumers is P_c. The difference in these two prices is $1, which is the amount per unit received by the taxing authority.[6] Note that, despite the fact that the tax is levied on consumers, both consumers and producers bear the burden of the tax. Consumers incur a reduction in consumer surplus because they are now paying a higher price per unit for a reduced quantity. Meanwhile, producers incur a reduction in producer surplus because they are now receiving a lower price per unit and producing less.

Does it matter in this case if producers had been taxed $1 per unit of output rather than taxing consumers $1 for every unit they purchased? The answer is no. This possibly surprising result is shown in Figure 5. Because the tax increases production costs by $1 per unit, the supply curve is shifted upward by the size of the tax. Once again, the difference between the price paid by consumers, P_c, and the net price received by producers, P_p, is $1. This tax wedge is identical to

[6] To distinguish between the two prices, the price paid by consumers can be called the gross-of-tax price, while the price received by producers can be called the net-of-tax price.

Figure 5

Effects of a Tax Levied on Producers

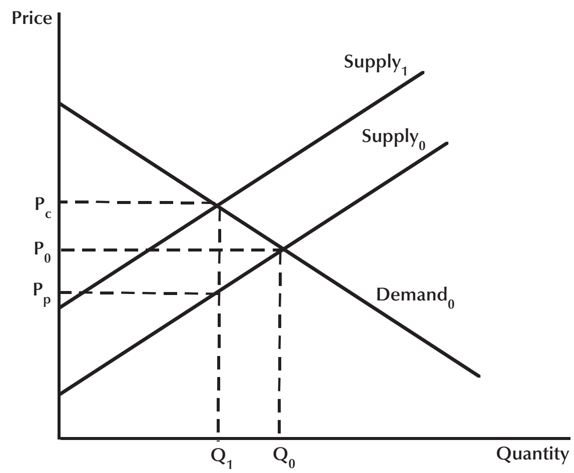
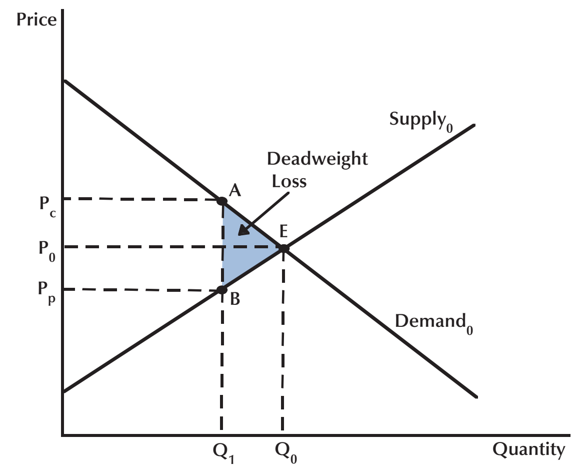


Figure 6

Deadweight Loss of a Tax



the tax wedge shown in Figure 4. Consequently, the equilibrium quantity, Q_1, in Figure 5 is the same quantity as in Figure 4. Given the identical price faced by consumers and the identical quantity, consumers bear the same burden, regardless of upon whom the tax is levied. An identical comment can be made concerning the tax burden of producers.

Taxes and Welfare

Next, let's examine the welfare consequences in more detail. Because Figures 4 and 5 generate identical results, we can simplify the analysis by not showing the curve that shifts. Figure 6 shows the same information as Figures 4 and 5. The $1 tax per unit drives a $1 wedge between what consumers pay and producers receive. This $1 per unit is received by the government as tax revenue. In Figure 6, tax revenue is represented by the rectangle P_pP_cAB (or $(P_c - P_p) \times Q_1$). Meanwhile, the tax imposes burdens on consumers and producers. The cost of taxation for consumers and producers reflects not only the amount paid to the taxing authority, but also the cost associated with transactions that no longer occur because the tax has made them "uneconomic." This latter point is due to the fact that, in nearly all cases, taxation causes consumers and producers to change their behavior.

The tax burden for consumers is the reduction in consumer surplus. Due to the higher price and the reduced consumption, this loss is the four-sided area P_0P_cAE. The tax burden for producers is the reduction in producer surplus. Due to the lower price and the reduced production, this loss for producers is the four-sided area P_pP_0EB. Note that some of the losses incurred by both consumers and producers reflect a transfer to the government.[7] Overall, the net decline in national well-being is the triangle BAE, which is termed the deadweight loss (DWL) caused by the tax. This loss reflects the fact that taxation prevents some mutually beneficial exchanges between consumers and producers from occurring.

For a given tax, the distribution of the effects on economic well-being and the size of the DWL depend on how much quantity demanded and quantity supplied respond to the change in price stemming from the tax. A summary measure of this responsiveness is the price elasticity of demand (supply). The price elasticity of demand is the absolute value of the percentage change in

[7] To simplify the analysis, we have assumed that the government programs funded by the tax revenues yield benefits that are equal to the losses incurred by consumers and producers associated with the tax payments.

quantity demanded divided by the percentage change in price. Larger values of the price elasticity of demand are associated with flatter slopes of the demand curve; conversely, smaller values of the price elasticity of demand are associated with steeper slopes of the demand curve.[8,9] The same terminology and implications for slope are used for the price elasticity of supply, for which larger (smaller) values of the price elasticity of supply are associated with flatter (steeper) slopes of the supply curve.

The price elasticities of demand and supply affect the tax burdens imposed on consumers and producers. Exactly how is straightforward. Begin by envisioning a clockwise rotation of the demand curve around point E in Figure 6. In this case, the price elasticity of demand is becoming less elastic (more inelastic). With an unchanged supply curve, as the demand curve becomes less elastic, the price that consumers pay will rise as will the price that producers receive. Note that the tax wedge must remain constant. The end result is that the tax burden imposed on consumers rises relative to that of producers. In other words, holding all other things constant, as the price elasticity of demand decreases, the tax burden of consumers rises relative to that of producers.

Similar results occur when the price elasticity of supply becomes less elastic. With the demand curve unchanged, as the supply curve becomes less elastic, the price that producers will receive falls as will the price that consumers pay. Thus, the tax burden imposed on producers rises relative to that of consumers. In summary, as the price elasticity of demand (supply) decreases, the larger the relative tax burden of consumers (producers).

The economic intuition underlying this result is straightforward. In terms of their consumption, the less responsive consumers are to the higher price they pay as a result of a tax, the higher their relative tax burden. In other words, the tax burden of consumers is relatively more the less they change their behavior in response to higher prices. Similar reasoning pertains to producers. In terms of their production, the less responsive producers are to the lower (net) price they receive as a result of a tax, the higher their relative tax burden.

In addition to being related to the tax burdens of consumers and producers, the price elasticities of demand and supply affect the DWL of the tax. A tax creates a DWL because it causes buyers and sellers to change their consumption and production behavior. For a given tax, the larger the price elasticities of demand and supply, the larger are the changes in consumption and production. Thus, larger price elasticities of demand and supply are associated with larger DWLs.

THE THEORY OF LAND TAXATION

> "Tax something, there will be less of it—except land." (Harriss, 2003)

Land is different from most other goods. Namely, proponents of land taxation note that the supply of (unimproved) land, which is provided by nature, is fixed. In other words, the supply is perfectly inelastic. This implies that the supply curve for land is vertical, as shown in Figure 7. Recall that a tax on consumers of land (as well as a tax on consumers of any good, as outlined in the previous section) will shift the demand curve downward. Figure 7A shows a decline in demand from Demand_0 to Demand_1. Because the supply curve is vertical, shifting the demand curve downward implies that the new demand curve will intersect the supply curve at the same quantity of land as before the land tax (i.e., Q_0). As a result, the intersection of the new demand curve and supply curve will occur at a lower net-of-tax equilibrium price than before the land tax (i.e., P_1 rather than P_0)—with no change in the equilibrium quantity of land. Further, note that the gross-of-tax price, P_0, is identical to the price in the absence of the tax. The land tax has no effect on the allocation of productive resources. The

8 The price elasticity of demand at a specific point on a demand curve, for instance (P_0,Q_0), is equal to the absolute value of P_0/Q_0 times $\Delta Q/\Delta P$. The latter term is the inverse of the slope of the demand curve. If a demand curve were to become flatter (steeper), the value of $\Delta Q/\Delta P$ would increase (decrease). Thus, the price elasticity of demand increases (decreases).

9 Economists refer to increasing values of the price elasticity of demand (supply) as being "more elastic" or "less inelastic" and decreasing values as being "less elastic" or "more inelastic."

result is that land owners bear the entire burden of the tax and there is no DWL with this land tax.[10]

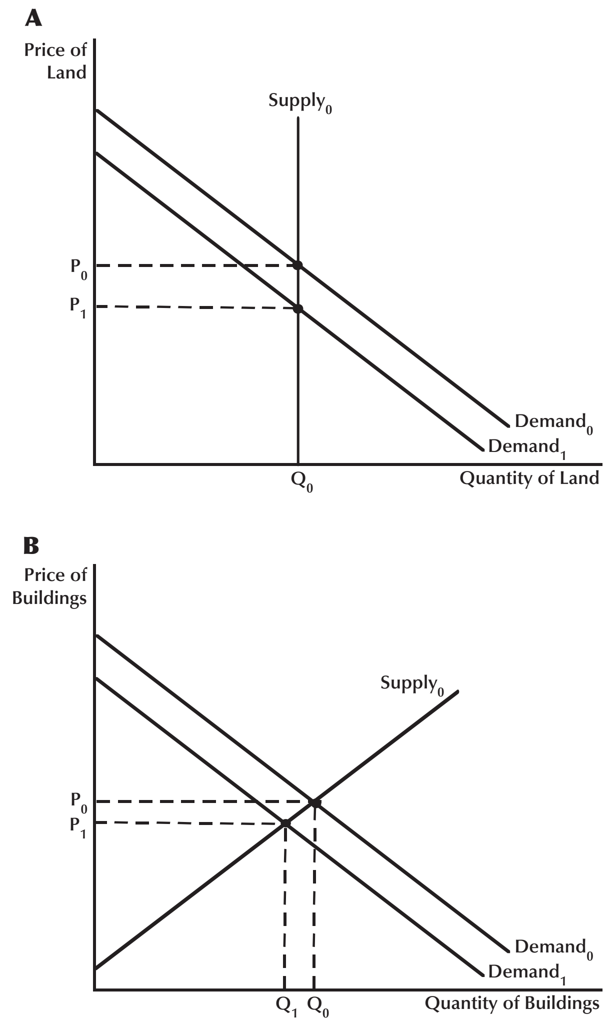
A property tax on buildings, however, alters or distorts behavior away from that which would take place in a competitive economy without taxes. As seen in Figure 7B, because higher prices encourage producers to supply additional buildings, the supply curve for buildings slopes upward. The demand curve slopes downward for the same reason that the demand curve for land slopes downward—higher prices result in a decrease in the quantity of buildings demanded. A tax on consumers of buildings causes the demand curve to shift down by the amount of the tax. The end result is that some of the burden of the tax is borne by individuals who produce buildings (in the form of lower building prices) and fewer buildings are consumed in equilibrium. Thus, relative to the land tax case, the tax on buildings distorts behavior, leading to a DWL: The loss of consumer and producer surplus is greater than the revenue transferred to the government through the tax.

Moving from a traditional property tax (where land and buildings are taxed at one rate) to a two-rate tax (where land is taxed at a relatively higher rate) lowers DWL. If a goal is to keep total tax revenues unchanged, the tax on buildings can be lowered and the tax on land raised to achieve a revenue-neutral alternative. As a result, the distortionary (building) tax is decreased (i.e., in Figure 7B, the demand curve shifts *upward* from Demand$_1$), while the neutral (land) tax is increased. The overall effect is to lower the DWL. One way to think of this reduction in DWL is that a change to a more efficient tax system is equivalent to a tax cut. A given amount of public services can be provided with a lower local tax burden. Despite the fact that tax revenues are unchanged, the tax burden is effectively lowered because of the decline in DWL.

A reduction of the DWL associated with the current system of property taxation, however, is not the only economic argument that can be made to support increased tax rates on land and decreased tax rates on improvements. Mills (1998) has provided an analysis of the spatial equilibrium effects of single taxation. Specifically, Mills showed how productive activity throughout a hypothetical metropolitan area would be changed by a revenue-neutral switch from a conventional property tax (i.e., one that applies to both buildings [capital] and land) to a tax on land only. The switch causes capital and labor to be substituted for land. The more intensive use of capital and labor increases the productivity of each land

Figure 7

Taxing Land versus Taxing Buildings



[10] O'Sullivan (2003, pp. 526-28) presents a somewhat different exposition of the market effects of land taxation but arrives at the same outcome that the equilibrium price of land falls by the amount of the tax, with the quantity of land in equilibrium unchanged.

Table 1

Mill Rates: Kauai County, Hawaii, Fiscal Year 2004 (July 2003–June 2004)

	Homestead	Single family	Apartment	Hotel/ resort	Commercial	Industrial	Agricultural	Conservation
Buildings	3.64	4.5	8.15	8.15	8.15	8.15	4.5	4.5
Land	4.35	5.49	8.55	8.55	8.55	8.55	7.95	8.45

SOURCE: Tax Foundation of Hawaii; www.tfhawaii.org/taxes/property.html.

parcel, which tends to increase gross-of-tax land prices.[11] With agricultural land remaining untaxed, the city expands. Due to the increased use of capital and labor on each land parcel, output in the metropolitan area expands—in fact, output increases at every location within the metropolitan area. Thus, as shown by Brueckner (2001), such a tax change would lead to denser patterns of land development and, therefore, inhibit metropolitan sprawl. Since land in many inner cities is currently underutilized, such denser land development would be desirable.

In a subsequent section of our paper, we take a closer look at the possible magnitudes of the economic effects of using two-rate taxation. A caveat mentioned by Mills is that the adjustment period associated with such a tax change might well be very long and require very large, albeit justified, investments.

TWO-RATE TAXATION IN THE UNITED STATES

Because real property taxes in the United States are generally levied by local taxing authorities, most examples of two-rate taxation are at the local level. The most frequently mentioned example involves the state of Pennsylvania and, specifically, the city of Pittsburgh. In 1979-80, Pittsburgh revamped its property tax system by raising tax rates on land to more than five times the rate on structures. From 1913 to 1979-80, Pittsburgh had a two-rate tax in place, with the tax rate on buildings being twice the rate on land (Hartzok, 1997).[12] According to the state's *Taxation Manual* (Pennsylvania Department of Community and Economic Development, 2002), Scranton is the other city in Pennsylvania that is authorized to charge lower tax rates on buildings than on land.[13]

Although recent state legislation authorized two local governments in Virginia to implement a two-rate tax, neither has adopted it (Brunori, 2004). The cities of Fairfax and Roanoke have the authority to tax property on land at a lower rate than the corresponding land; however, one stipulation of the legislation is that the tax on property may not be zero, thus precluding a pure land tax.[14]

A final example involves Hawaii, whose state legislature passed a two-rate tax in 1963. A major difference between the Hawaii legislation and the Pittsburgh legislation, however, is that in Hawaii the two-rate tax applied to all jurisdictions (Rybeck, 2000). Legislation in 1978 granted counties the authority to set their own local property tax rates. As of fiscal year 2004, though, Kauai was the only one of Hawaii's four counties to set

[11] Technically speaking, the equilibrium gross-of-tax rent-distance function rises, while the net-of-tax rent-distance likely falls.

[12] The *Taxation Manual* for Pennsylvania (2002) notes that all property owners in the state pay property taxes to the municipality, the corresponding county, and school district. Oates and Schwab (1997) also note that for the city of Pittsburgh, the presence of these school district and county property taxes that levy a "conventional property tax" imply that the net tax rate for land is greater than twice the rate for structures.

[13] Other governmental entities in Pennsylvania, such as third-class cities, boroughs, and third-class school districts coterminous with third-class cities, may also use two-rate taxation. For a discussion of the structure of different local levels of Pennsylvania government, see section 6 of volume 116 of the *Pennsylvania Manual* (2003): www.dgs.state.pa.us/pamanual/lib/pamanual/sec6/section6a.pdf.

[14] For a reproduction of the detailed Virginia legislation, see www.progress.org/cg/roan03.htm.

higher tax rates for land than for buildings.[15] As can be seen in Table 1, the differences in tax rates for many property classes for Kauai are small.

EFFECTS OF TWO-RATE TAXATION

A number of recent quantitative studies have examined the effects of implementing two-rate taxation. Quantitative studies are important for providing policymakers with estimates of the potential gains from shifting toward land as a tax base and some sense of the size of the distributional issues of such a change. To date, only one comprehensive study, focused on the consequences for Pittsburgh, has attempted to identify the actual effects of two-rate taxation. Generally speaking, existing quantitative studies have attempted to gauge the hypothetical effects of two-rate taxation. Some of these latter studies have focused on the short-run initial effects, while others have attempted to identify the likely effects given the economic decisions that would ensue under the changed tax regime. In other words, the former studies consider only the initial redistribution of the property tax burden and, as a result, do not identify how differential rates on land and improvements are likely to induce more intensive use of land.

The Pittsburgh Experience

After Pittsburgh further increased the difference between the tax rate for land and the tax rate for buildings in 1979-80, the city experienced a substantial increase in building activity. Oates and Schwab (1997) provided suggestive evidence that Pittsburgh's change in tax rates played a major role in stimulating the building boom.[16] They noted that this finding is surprising, however, because public finance theory suggests that increasing the land tax while leaving the buildings tax unchanged should have no effect on building activity. It is also worth noting, however, that according to Oates and Schwab, the city of Pittsburgh granted tax cuts for new building construction. These tax cuts essentially indirectly lowered the tax rate on new (but not on existing) buildings.

Table 2

Percent Change in Average Annual Value of Building Permits Between 1960-79 and 1980-89

City	Percent change
Akron	–34.4
Allentown	–40.2
Buffalo	–11.5
Canton	–39.7
Cincinnati	–27.2
Cleveland	–31.8
Columbus	15.4
Dayton	–14.4
Detroit	–24.7
Erie	–52.9
Pittsburgh	70.4
Rochester	–30.6
Syracuse	–43.2
Toledo	–32.4
Youngstown	–67.0

SOURCE: Oates and Schwab (1997, Table 3).

Suggestive evidence concerning the impact of Pittsburgh's change in tax rates is presented in Table 2, which shows the percentage change in average annual value of building permits for 15 cities between the two periods 1960-79 and 1980-89. Excluding Pittsburgh, which had a greater than 70 percent increase in average annual building permits between these two periods, and Columbus, which had a 15 percent increase, all 13 other cities experienced a decrease in building permits between these two periods.

To generate stronger evidence, Oates and Schwab also used an econometric model to test the impact of structural change that may have occurred in 1979-80 when the city of Pittsburgh

15 For the details of property tax rates for all four counties in Hawaii, see www.tfhawaii.org/taxes/property.html.

16 Oates and Schwab also point out, however, that there was a major Pittsburgh revitalization effort (Renaissance II) underway in the late 1970s in an attempt to counteract the demise of the city's steel industry.

increased the land/buildings tax differential. For each of the 15 cities listed in Table 2, Oates and Schwab ran regressions with the real value of building permits against a constant and a dummy variable. This dummy variable took a value of zero for years prior to 1980 and a value of 1 for 1980 and later. They found that the coefficient on the dummy variable was both positive and statistically significant only in the regression for the city of Pittsburgh.[17]

To present further evidence on the impact of increasing the land/buildings tax rate differential, Oates and Schwab looked at U.S. Census Bureau data of the metropolitan statistical areas (MSAs) of these 15 cities over the years 1974-89. With these data, they made a distinction between cities and suburbs and between the real values of residential and nonresidential building permits. They ran regressions with these Census data using the same specifications as for the Dun and Bradstreet data, with distinct regressions for residential, nonresidential, and office building permits for each city and its respective suburbs. For the city of Pittsburgh, the post-1979 dummy was significant in the regressions for both the nonresidential and office building permits, but insignificant for the residential regression.[18] For the Pittsburgh suburb regressions, the post-1979 dummy in the residential permit regression was significant but negative; this dummy was significant but positive in the Pittsburgh suburban office regression, and insignificant in the Pittsburgh suburban nonresidential regression.[19]

These findings reveal a correlation between the 1979-80 tax reforms in Pittsburgh and the subsequent increases in building permits. As mentioned above, however, these findings are far from definitive in light of public finance theory and the specifics of the tax reform in Pittsburgh (that is, the fact that the tax on buildings was unchanged when the tax on land was raised).[20]

Short-Run Initial Effects

Because Virginia is a state whose citizens have shown much interest in two-rate taxation, Bowman and Bell (2004) used data on individual property parcels from three Virginia locations to estimate property tax liabilities using a tax in which only the unimproved land is taxed. Consequently, Bowman and Bell were able to identify the initial change in the real property tax liabilities of taxpayers resulting from a shift to a pure land tax, a limiting case of two-rate taxation, from the current uniform tax on land and improvements. The three areas examined vary substantially from each other. Roanoke, a city of roughly 100,000 residents, has been slowly losing population, yet has experienced job growth. Chesterfield County, a bedroom county in the Richmond area with over 250,000 residents, has grown both in terms of population and jobs. Highland County, a small rural county of less than 2,500 residents, has experienced declines in both population and jobs.

Regardless of the area, Bowman and Bell found that owners of properties with high land-to-improvements ratios will tend to experience an increase in their tax liabilities with the move to two-rate taxation, while owners of properties with low land-to-improvements ratios will tend to experience a decrease in their tax liabilities. Generally speaking, owners of residential property, especially owners of multi-unit housing properties, would tend to benefit. In addition, the researchers found that, even within a specific classification of land use, substantial differences in distributional effects were likely.

Simulation Studies

The limited use of two-rate taxation has motivated a number of informative simulation studies

[17] Oates and Schwab also looked at a slight variation of the aforementioned model, which included a time trend in addition to the other variables. They found that the coefficient on the post-1979 dummy variable was positive and statistically significant only in the Pittsburgh and Buffalo regressions.

[18] The signs and significance of these variables are the same for the similar specifications that include time trends.

[19] Once again, the signs and significance of these variables are robust to the inclusion of a time trend in these regressions.

[20] The Pittsburgh City Council removed the two-rate system in 2000. Craig (2003) reports that construction spending in the city was higher in the two years prior to rescission than the two years after. Construction activity in the city was also lower than in the surrounding suburbs and in the United States as a whole after the rescission.

to generate information about its potential effects. It is important to stress, similar to Kodrzycki (1998), some important caveats about simulation studies. Computable general equilibrium models, which provide the foundation for simulation studies, provide a range of answers to policy questions because the appropriate structure of the underlying model and the choice of parameter values are subject to much uncertainty.[21] Here we focus on three studies. These studies provide insights into the economic consequences of land value and two-rate taxation in a variety of situations.

Nechyba (2001) has explored the economic impacts of land tax reforms for each U.S. state as well as for an average state. Numerous revenue-neutral reforms were examined; in other words, the increase in revenues from increased taxes on unimproved land exactly matches the decrease in tax revenues from reducing some distortionary tax on capital or labor. Generally speaking, based on the likely change in land prices, Nechyba found that reforms eliminating entire classes of taxes are feasible in nearly all states. The political prospects for passing a specific reform are better in states with high per capita taxes and low per capita incomes and when the reform is targeted to lowering taxes on capital rather than labor. In addition, reforms targeted to lowering taxes on capital cause either increases in land prices or modest declines, while reforms targeted to lowering taxes on labor tend to cause large declines in land prices.

The second study we examine was done by England (2003). He undertook a simulation study using county-level data that examined a revenue-neutral shift for New Hampshire from a uniform property tax to a land value tax. The shift of the tax burden from capital to land reduces the disposable income of owners of land, some of which is likely borne by nonresidents. The reduction in disposable income leads to reduced consumer spending on items other than housing services. These impacts, however, are more than offset by the changes set in motion by the decline in the cost of owning residential buildings as well as commercial and industrial capital. As a result, residential construction and business investment spending are boosted. Overall, employment and gross state product increase in New Hampshire, both immediately and after ten years. Moreover, each of New Hampshire's ten counties is projected to have higher output, income, employment, and population a decade after the tax change. While all counties benefit, the economic changes for the county that benefits the most are roughly double those of the county benefiting the least.

The final study we examine was done by Haughwout (2004), who estimates the consequences for New York City of replacing its current tax system with a land tax. Two situations are examined. In one case all taxes are eliminated with the exception of the land tax, which is maintained at its current rate. In the other case, the key difference is that the tax rate on land is increased so that total tax revenue is maintained.

In the first case, the distortions caused by taxes are eliminated and overall tax burdens are reduced. At the same time, tax revenues decline, so the provision of public goods falls correspondingly. Overall, New York City experiences substantial increases in private output, private capital stock, employment, land values, and population and a substantial reduction in public good provision and per capita tax revenues.

In the second case, the tax rate on land is increased substantially so that tax revenues are maintained. Contrary to the first case, land prices fall, due in part to the substantially higher land tax rate. Public goods provision is maintained. Similar to the first case, private output, private capital stock, employment, and population rise sharply.

The two cases examined by Haughwout produce a clear message. The potential gains from eliminating the distortions stemming from the taxation of capital and labor, especially in a city in which existing tax rates are relatively high,

[21] For example, Kodrzycki (1998) highlights the importance of the elasticity of substitution between land and capital, which is the optimal response of the capital/land ratio to a change in the relative prices of these inputs. When this elasticity equals 0.25, Nechyba (1998) finds that the substitution of land taxes for capital taxes leads to an increase of 43 percent in the capital/land ratio and 32 percent in national output. Meanwhile, a higher value of this elasticity, 0.5, is associated with a more than doubling of the capital/land ratio and an 89 percent increase in national output.

are quite large. Such a conclusion leads quite naturally to the issue of why two-rate taxation is rarely used.

IMPLEMENTING TWO-RATE TAXATION: SOME PRACTICAL PROBLEMS

In opposition to the theory and evidence of the potential gains from using two-rate taxation, a number of practical problems face a region that decides to implement two-rate taxation. These problems include valuing land accurately, determining the revenue potential of land value taxation, and providing sufficient public infrastructure to support the increased economic activity.

To impose different tax rates on improvements to land and raw land, one must have estimates of the value of the improvements and the value of raw land. Netzer (1998), using an in-depth examination of land value data, concluded that nonagricultural land values could not be trusted. Moreover, the data were incomplete with respect to timing and coverage. Therefore, practically speaking, useful land value data for two-rate taxation purposes do not exist.

As Mills (1998) has stressed, to preclude distortions, a land tax must be applied to the value of land prior to any improvements. Defining exactly what raw land is presents problems. When a parcel of land is ready to be developed, it has already been improved substantially. Preparing land for development generally requires a number of costly activities, such as clearing and leveling the land, conducting environmental tests, surveying, obtaining the required permits, and installing underground infrastructure. Furthermore, the value of raw land hinges on the state of technology as well as on the state of urban and rural development. For example, agricultural inventions have affected the value of rural land, while construction innovations have affected the value of urban land. It remains to be seen how developments in information technology will affect land values. The bottom line is that estimating the value of raw land, which is likely to change over time, is very challenging.

Even more challenging is the assessment of land values of developed properties. With respect to commercial property, Mills (1998) notes the sites and the structures are owned by different groups. The separate ownership is frequently driven by tax considerations, with the site owned by an untaxed organization and the structure owned by a business in a high tax bracket that can utilize the benefits of depreciation. In theory, separate estimates of site and structure values of developed properties could be generated using an approach known as hedonic pricing. Such an approach is commonly used to explain, in a statistical sense, housing prices. The sales price of a house is related to the characteristics of the house (i.e., living space, number of bathrooms, age, etc.), lot size, the neighborhood, and the community.

Applying hedonic pricing to commercial properties is problematic. Difficulties would arise because of a lack of agreement as to which characteristics should be included, the paucity of transactions, and the fact that many transactions are not arms-length exchanges. Consequently, generating accurate estimates of raw land values, an essential component of land taxation, is difficult; uniform taxation may be preferred because it is less costly to use than two-rate taxation.[22] However, whether the additional administrative cost is large or small is unclear. Netzer (1998) has noted that, despite the fact that each parcel of land is unique, the difference in value for adjacent parcels is minimal, a fact that should ease the administrative burden.

The absence of accurate land value data makes it difficult to answer the question of whether land value taxation would generate sufficient revenues to be an important replacement for revenues from conventional property taxes. Despite the lack of land value data, Netzer (1998) and McGuire (1998) find that Pittsburgh's experience with two-rate taxation suggests that land value taxation can generate an adequate level of tax revenue. On the other hand, Mills (1998) is doubtful. His reasoning is straightforward: Annual real estate taxes are 1.5 to 2.0 percent of the market value of tax-

[22] One way to overcome this problem, suggested by Anas (1998), is to have the city purchase and demolish some buildings and then sell the land.

able property and site values are estimated to be, at most, 10 percent of property values. If site and structure rents are capitalized at the same rate, gross-of-tax site rents are, at most, 1 percent of property values. Consequently, a 100 percent land rent tax would not generate sufficient revenue to replace the revenue from existing real estate taxes.

A land tax rate of 100 percent of the land rent is equivalent to land confiscation without assuming the liabilities of ownership. Mills (1998) has noted that courts have consistently ruled that similar regulations or taxes require, based on the Fifth Amendment, that owners be compensated for their losses. Such a court decision would negate the value of using the land tax. While it is unlikely that proponents of two-rate taxation would argue for a tax rate of 100 percent on land, there remains a question as to what percentage of land rent could be taxed away without substantially affecting an owner's incentive to seek the best use for the land. At some tax rate, major misallocations of land use would result.

Another potential problem occurs as a result of the increased economic activity that takes place, assuming the successful implementation of land taxation. The resulting increase in the building/land and employment/land ratios would necessitate increased infrastructure provided by government, such as transportation facilities and schools. Without transportation infrastructure, increased traffic congestion could negate the potential benefits of land taxation. The unanswered question is whether the political process would be responsive to the changed environment in the private sector. In light of the increased activity, many would downplay this situation as a problem, but rather view it as an opportunity.

THE POLITICAL ECONOMY OF TWO-RATE TAXATION

A strong case exists that two-rate taxation is more efficient than, and thus preferable to, uniform taxation. A reasonable question is why uniform taxation remains the norm. In addition to the practical problems discussed in the preceding section, a number of explanations have been proposed. These explanations fall into either of two general categories—one stressing that the efficiency gains are likely to be elusive and another stressing that opposition from those likely to be harmed by the change to a two-rate system prevents such a change.

Efficiency

We begin with a discussion of the arguments based on efficiency. Lee (2003) has shown that uniform taxation of land and capital may be more efficient than the taxation of land only. This possibility arises when some land in a taxing jurisdiction is owned by nonresidents. In terms of public policy, a specific jurisdiction is assumed to structure its fiscal policies in the interests of its residents. Thus, one might argue that tax policies are made with minimal consideration for the well-being of absentee owners because nonresidents do not vote in the jurisdiction. One consequence is that the jurisdiction taxes land excessively to exploit absentee owners and the resulting funds are used to overprovide public goods.[23] One way to mitigate the inefficiency of overtaxing land is for a higher-level government to require jurisdictions to tax land and capital at a uniform rate. This is what occurs in the United States because most state governments do not allow lower-level governments to deviate from uniform taxation.

Another argument suggesting the desirability of uniform taxation has been made by Wildasin and Wilson (1998). They start with the observation that the returns to land are risky under production uncertainty. This feature of the economy provides an incentive for individuals to diversify their risk by owning land in multiple jurisdictions. However, if each jurisdiction eliminates the rent on owning land with 100 percent tax rates on land, the benefits of diversification are eliminated.

[23] Public goods, according to Rosen (1995, p. 61), are goods characterized by "non-rival consumption." Nonrival consumption exists when one person's consumption of the good does not reduce its availability to anyone else. Common examples are national defense, lighthouses, roads, and parks. Note that for roads and parks, at some point, as more and more individuals attempt to enjoy the services of roads and parks congestion costs arise; when this occurs, consumption is no longer nonrival.

Therefore, uniform taxation allows for benefits from diversification and may be superior to the pure land tax.

Opposition

The argument that political opposition would mount against the change from uniform taxation to two-rate taxation is straightforward. The change in taxation will create winners and losers. Owners of properties with high land-to-improvements ratios (e.g., car dealerships) will tend to experience an increase in their tax liabilities with the move to two-rate taxation, while owners of properties with low land-to-improvements ratios (e.g., high-rise office buildings) will tend to experience a decrease in their tax liabilities. The owners of substantial amounts of land are likely to be wealthy and may have a disproportionate voice in the political process and, thus, prevent a change that would harm them.

The preceding discussion suggests that communities with heterogeneous consumer preferences and incomes might be unlikely candidates to adopt a land tax. On the other hand, more homogeneous areas, such as a suburban community, are more likely candidates for adoption. However, as Hamilton (1976) and Fischel (1998) have noted, these more homogeneous communities are also likely to be less afflicted by distortionary taxes.

A fundamental question concerns how two-rate taxation can be introduced so as to reduce the political opposition. England (2004) runs simulations using tax parcel data and shows that the opposition to tax reform will likely be reduced if, as part of the introduction of two-rate taxation, uniform property tax credits are also introduced.

Before completing our discussion concerning opposition to two-rate taxation, a few points about the knowledge of policymakers are warranted. A lack of understanding of two-rate taxation on the part of political leaders likely is not a reason for the limited use of the two-rate tax in states and localities in the United States. Brunori (2004) conducted a survey of state, county, and city officials and received about 1200 responses. The results indicate that between 65 and 70 percent of the respondents were "very or somewhat familiar" with land value taxation, and about 65 to 67 percent of these political leaders responded the same for the two-rate tax. About 76 percent of city and county government officials and over 62 percent of state lawmakers thought that a two-rate tax would enhance economic development. According to Brunori, over 40 percent of political leaders who responded held a common "misperception" that the two-rate tax system would lead to greater sprawl, due to additional building on undeveloped suburban land stemming from the reduction in the tax rate on structures.

CONCLUSION

Proponents of two-rate taxation stress that the taxation of real property involves two taxes. One falls on man-made capital, such as buildings, while the other falls on land, which is provided by nature. The taxation of capital tends to deter its formation. The higher the tax rate is in a specific location, the larger the incentive for investors to direct their capital elsewhere. The taxation of land, however, does not deter either the formation of land or encourage its relocation because land is essentially fixed in quantity and immobile. Therefore, the taxation of land does not generate the changes in behavior that one sees with the taxation of capital. This differential effect of taxation provides a justification for real property taxation that taxes buildings and land at different rates.

The theoretical gains associated with a revenue-neutral movement from single-rate taxation of real property to two-rate taxation are subject to little controversy. Gains arise in the form of declines in the deadweight losses associated with taxation and increases in overall economic activity. Parcels of land within a city would tend to be used more productively. However, the size of the gains associated with specific two-rate proposals is subject to much uncertainty. For example, the study of Pittsburgh's experience with two-rate taxation by Oates and Schwab (1997) and a number of simulation studies have suggested that the gains can be substantial. On the other hand, the paucity of experience with two-rate taxation, the sensitivity of the results of simulation studies to the underlying model's structure and the choice of parameter values, and concerns about administrative feasibility raise questions about the size of the gains that could be realized.

The majority of legislators are familiar with the theoretical consequences of two-rate taxation. However, the fact that few regions use two-rate taxation reflects the existence of significant obstacles. First, the practical implementation of two-rate taxation complicates the assessment process because the value of land must be separated from the value of improvements. Second, significant political opposition to two-rate taxation arises because the change to two-rate taxation causes some individuals to suffer adverse distributional consequences; generally speaking, owners of property with high land-to-improvement ratios tend to be harmed, while owners of property with low land-to-improvement ratios tend to benefit. In light of these consequences, policies that mitigate the adverse effects, yet allow for the capture of the economic gains, are required to reduce the opposition to two-rate taxation and increase the prospects for adoption.

Given the current system of taxation in the United States, pressures for using two-rate taxation will likely continue to emerge at the local level. It remains to be seen, however, whether the opinions of Nobel Prize winners Milton Friedman and William Vickrey concerning land taxation will become widely held by legislators and voters.

REFERENCES

Anas, Alex. "Commentary," in Dick Netzer, ed., *Land Value Taxation: Can It and Will It Work Today?* Cambridge, MA: Lincoln Institute of Land Policy, 1998, pp. 49-59.

Bowman, John H. and Bell, Michael E. "Implications of a Split-Rate Real Property Tax: An Initial Look at Three Virginia Local Government Areas." Working Paper WP04JB1, Lincoln Institute of Land Policy, 2004.

Bromley, Daniel W. "Commentary," in Dick Netzer, ed., *Land Value Taxation: Can It and Will It Work Today?* Cambridge, MA: Lincoln Institute of Land Policy, 1998, pp. 24-30.

Brueckner, Jan K. "Property Taxation and Urban Sprawl," in Wallace E. Oates, ed., *Property Taxation and Local Government Finance: Essays in Honor of C. Lowell Harriss*. Cambridge, MA: Lincoln Institute of Land Policy, 2001, pp. 153-72.

Brunori, David. "What Politicians Know About Land Taxation." *Land Lines*, October 2004, *16*(4); www.lincolninst.edu/pubs/pub-detail.asp?id=972.

Craig, Eleanor D. "Land Value Taxes and Wilmington, Delaware: A Case Study," Proceedings of the National Tax Association's 96th Annual Conference on Taxation, November 14, 2003, pp. 275-78; www.buec.udel.edu/craige/nta-lvt.htm.

England, Richard W. "State and Local Impacts of a Revenue-Neutral Shift from a Uniform Property to a Land Value Tax: Results of a Simulation Study." *Land Economics*, February 2003, *79*(1), pp. 38-43.

England, Richard W. "An Essay on the Political Economy of Two-Rate Property Taxation." Working Paper WP04RE1, Lincoln Institute of Land Policy, 2004.

Fischel, William A. "The Ethics of Land Value Taxation Revisited: Has the Millennium Arrived Without Anyone Noticing?" in Dick Netzer, ed., *Land Value Taxation: Can It and Will It Work Today?* Cambridge, MA: Lincoln Institute of Land Policy, 1998, pp. 1-23.

George, Henry. *Progress and Poverty*. New York: Schalkenbach Foundation, [1879] 1954.

Hamilton, Bruce W. "Capitalization of Intrajurisdictional Differences in Local Tax Prices." *American Economic Review*, December 1976, *66*(5), pp. 743-53.

Harriss, C. Lowell. "Taxing for Progress." Unpublished manuscript, June 2003.

Hartzok, Alanna. "Pennsylvania's Success with Local Property Tax Reform: The Split Rate Tax." *The American Journal of Economics and Sociology*, April 1997, *56*(2), pp. 205-13.

Haughwout, Andrew F. "Land Taxation in New York City: A General Equilibrium Analysis," in Amy Ellen Schwartz, ed., *City Taxes, City Spending: Essays in Honor of Dick Netzer*. London: Edward Elgar, 2004.

Kodrzycki, Yolanda K. "Replacing Capital Taxes with Land Taxes: Efficiency and Distributional Implications with an Application to the United States Economy: Commentary," in Dick Netzer, ed., *Land Value Taxation: Can It and Will It Work Today?* Cambridge, MA: Lincoln Institute of Land Policy, 1998, pp. 205-09.

Lee, Kangoh. "Should Land and Capital Be Taxed at a Uniform Rate?" *Canadian Journal of Economics*, May 2003, *36*(2), pp. 350-72.

Mankiw, N. Gregory. *Principles of Economics*. Mason, OH: Thompson/South-Western, 2004.

McGuire, Therese J. "The Relevance and Feasibility of Land Value Taxation in the Rich Countries: Commentary," in Dick Netzer, ed., *Land Value Taxation: Can It and Will It Work Today?* Cambridge, MA: Lincoln Institute of Land Policy, 1998, pp. 137-40.

Mills, Edwin S. "The Economic Consequences of a Land Tax," in Dick Netzer, ed., *Land Value Taxation: Can It and Will It Work Today?* Cambridge, MA: Lincoln Institute of Land Policy, 1998, pp. 31-48.

Nechyba, Thomas. "Replacing Capital Taxes with Land Taxes: Efficiency and Distributional Implications with an Application to the United States Economy," in Dick Netzer, ed., *Land Value Taxation: Can It and Will It Work Today?* Cambridge, MA: Lincoln Institute of Land Policy, 1998, pp. 183-204.

Nechyba, Thomas. "Prospects for Land Rent Taxes in State and Local Tax Reform." Working paper, Lincoln Institute of Land Policy, 2001.

Netzer, Dick. "The Relevance and Feasibility of Land Value Taxation in the Rich Countries," in Dick Netzer, ed., *Land Value Taxation: Can It and Will It Work Today?* Cambridge, MA: Lincoln Institute of Land Policy, 1998, pp. 109-36.

Oates, Wallace E. and Schwab, Robert M. "The Impact of Urban Land Taxation: The Pittsburgh Experience." *National Tax Journal*, March 1997, *50*(1), pp. 1-21.

O'Sullivan, Arthur. *Urban Economics*. Fifth Edition. New York: McGraw-Hill/Irwin, 2003.

Pennsylvania Department of Community and Economic Development, Commonwealth of Pennsylvania. *Taxation Manual*. Eighth Edition. October 2002; www.inventpa.com/docs/Document/application/pdf/9ab7d259-e9a3-4340-af5b-a5deed9757aa/taxation.pdf.

Pennsylvania Manual, Commonwealth of Pennsylvania. Section 6: Local Government. Volume 116, 2003; www.dgs.state.pa.us/pamanual/lib/pamanual/sec6/section6a.pdf.

Rosen, Harvey S. *Public Finance*. Fourth Edition. Chicago: Irwin, 1995.

Rybeck, Walter. "United States," in Robert V. Andelson, ed., *Land Value Taxation Around the World*. Chap. 9. Malden, MA: Blackwell Publishers, 2000.

Vickrey, William. "Simplifications, Progression, and a Level Playing Field," in Kenneth Wenzer, ed., *Land Value Taxation: The Equitable and Efficient Source of Public Finance*. Armonk, NY: M.E. Sharpe, 1999, pp. 17-23.

Wildasin, David E. and Wilson, John Douglas. "Risky Local Tax Bases: Risk-Pooling vs. Rent-Capture." *Journal of Public Economics*, August 1998, *69*(2), pp. 229-47.

Evidence on Wage Inequality, Worker Education, and Technology

Christopher H. Wheeler

The rise in U.S. wage inequality over the past two decades is commonly associated with an increase in the use of "skill-biased" technologies (e.g., computer equipment) in the workplace, yet relatively few studies have attempted to measure the direct link between the two. This paper explores the relationship among inequality, worker education levels, and workplace computer usage using a sample of 230 U.S. industries between 1983 and 2002. The results generate two primary conclusions: First, this rising inequality in the United States has been caused predominantly by increasing wage dispersion within industries rather than between industries. Second, within-industry inequality is strongly tied to both the frequency of computer usage among workers and the fraction of total employment with a college degree. Both results lend support to the idea that skill-biased technological change has been an important element in the rise of U.S. wage inequality.

Federal Reserve Bank of St. Louis *Review*, May/June 2005, *87*(3), pp. 375-93.

The rapid rise of U.S. wage inequality in recent decades has produced a sizable literature both documenting the empirical trends and theorizing about their causes.[1] The main empirical findings can be summarized by three basic patterns. First, the overall distribution of hourly and weekly earnings across all workers in the economy has grown wider. Second, consistent with this rise, the wage gaps between workers with different levels of education, especially between college graduates and workers with no more than a high school diploma, have also increased. This rise in "between-education-group" earnings disparity, however, accounts for only a modest fraction of the rise in overall wage dispersion because of the third pattern: The variance of wages among workers with the *same* level of education has also grown.[2]

To explain these patterns, a variety of theories have been advanced, including those stressing the growth of international trade, changes in institutions (e.g., declining unionization and real minimum wage), rising immigration, and technological change. Growing levels of imports into the United States, for instance, may have hit workers in trade-sensitive industries (e.g., textiles and apparel) particularly hard as domestic labor demand and, consequently, wages have dropped.[3] Rising immigration since the 1960s may also have contributed to these trends by increasing the supply of less-skilled workers in the U.S. labor market (Borjas, Freeman, and Katz, 1997). In addition, because unionization is often associated with wage compression (Fortin and Lemieux, 1997), declining rates of union membership in the United States may have contributed to rising earnings disparity.[4]

[1] See Levy and Murnane (1992) and Acemoglu (2002) for surveys.

[2] These basic patterns have also been observed for a number of Organisation for Economic Co-operation and Development (OECD) countries, although wage inequality in the United States remains higher than that of most other developed economies. See Blau and Kahn (1996).

[3] The Bureau of Labor Statistics produces the *Occupational Outlook Handbook*, which offers predictions about job and wage growth in various sectors, including Textiles and Apparel, given (among other things) trends in international trade. The most recent edition can be found at www.bls.gov/oco/home.htm.

[4] In addition to Fortin and Lemieux (1997), Topel (1997) and Johnson (1997) offer surveys of several prominent theories of wage inequality.

Christopher H. Wheeler is an economist at the Federal Reserve Bank of St. Louis. Elizabeth La Jeunesse provided research assistance.

While there remains some disagreement as to the significance of each of these mechanisms, a general consensus has formed around one particular theory: skill-biased technological change (SBTC).[5] The hypothesis is quite simple. Over the past several decades, the supply of highly educated workers in total employment has grown. In 1950, for example, 17 percent of U.S. workers had some education at the college level. By 1990, 57 percent did.[6] As a result of this increase, the return to investing in technologies that complement the skills of these highly educated workers—the most commonly cited example of which is information technology—also rose because the search costs involved in finding and hiring "skilled" labor declined.[7] Accordingly, recent technological change has served to boost the wages of skilled workers while depressing the employment opportunities and earnings of the less-skilled. As noted by Acemoglu (1999), if "skills" are positively but imperfectly associated with educational attainment, this mechanism can lead to larger between-education-group gaps as well as greater inequality *within* education groups.

Although there has been a host of evidence documented on the general topic of technological change, skill distributions, and inequality, there are (at least) two issues that remain unresolved. First, to what extent is rising inequality a within- or between-sector phenomenon? That is, does the SBTC argument imply that inequality is driven by growing differentials across workers in different sectors or by growing gaps within the same sector? Caselli (1999), for example, reports that, between 1975 and 1991, the variance of (equipment) capital-labor ratios across 450 four-digit manufacturing sectors rose sharply in the United States. Because capital-labor ratios tend to correlate positively with both wages and the fraction of highly skilled workers in total employment, Caselli interprets this trend as evidence that SBTC has been highly variable across sectors: Some industries have adopted advanced technologies and hired educated workers; others have chosen to utilize skill-*un*biased technologies and less-educated workers. This result suggests that SBTC has driven inequality higher through a between-industry channel. A qualitatively similar result is reported by Acemoglu (1999), who finds that the fraction of workers in the United States holding jobs (defined by 174 industry-occupation cells) in the bottom and top tails of the distribution of average hourly pay increased between 1983 and 1993. This finding, he concludes, indicates that workers have increasingly been sorted into "good" and "bad" jobs, which further suggests that rising inequality has been the product of growing between-sector dispersion.

On the other hand, many studies of inequality (e.g., Katz and Murphy, 1992; Juhn, Murphy, and Pierce, 1993; Card and DiNardo, 2002) suggest that, even after accounting for observable differences across workers (including their industries of employment), the dispersion in their wage earnings has risen markedly. Such evidence suggests that within-sector differences must also be an important aspect of rising dispersion. However, it remains unclear just how important these two elements have been in explaining the overall increase in earnings disparity.

Second, much of the evidence on technological change and wage dispersion tends to be indirect. That is, in spite of the popularity of the SBTC hypothesis, surprisingly little research has directly examined the association between inequality and the extent to which computer equipment (or any other "advanced" technology) is used in production.[8] Most studies have either connected average wage earnings to the use of computers and other sophisticated technologies (Krueger, 1993; Doms, Dunne, and Troske, 1997) or explored the relationship between the adoption of information technology and the distribution of worker skill (Berman, Bound, and Griliches, 1994; Doms, Dunne, and Troske, 1997; Autor, Katz, and Krueger, 1998).

5 As noted by Card and DiNardo (2002), the SBTC explanation is far from complete. Nonetheless, as suggested by Johnson (1997), there is wide agreement that it has been an important determinant of rising inequality.

6 These statistics are reported by Wheeler (2004).

7 See Autor, Levy, and Frank (2003) for an empirical analysis of why computers are considered skill-biased.

8 Dunne, Foster, and Troske (2004) is a notable exception. However, the focus of that paper is the U.S. manufacturing sector rather than the entire private U.S. economy, which I examine here.

This paper seeks to address both of these issues. To this end, I perform two exercises. In the first, I use annual data from the Current Population Survey (CPS) over the period 1983-2002 to evaluate the degree to which rising wage dispersion in the entire private U.S. economy can be attributed to growing dispersion within industries as opposed to between them. In the second, I look at the relationship between, on the one hand, a variety of inequality measures within individual industries and, on the other, the distribution of educational attainment and the extent of computer usage among workers employed in those industries. Computer usage, I assume, provides a direct measure of SBTC; educational attainment provides an indirect measure.[9]

To summarize briefly, the results indicate that the rising U.S. wage inequality has been driven primarily by growing dispersion among workers within the same industry rather than between industries. In each year, more than 75 percent of the variance of hourly earnings can be attributed to within-industry variation. More importantly, this fraction has grown steadily over time, suggesting that the majority of the rise in overall wage variance is due to increasing wage disparity among workers within the same industry. When I turn to the analysis of inequality within industries, I find that wage dispersion—measured in a variety of ways—is positively associated with both the fraction of college-educated workers and the extent of computer usage. These results, I conclude, offer some support for the skill-biased technological change argument.

DATA

Sources

The majority of the worker-level data used in this paper are derived from the Merged Outgoing Rotation Group (MORG) files of the CPS for each year between 1983 and 2002. These files are constructed by combining the individuals from each month's CPS who are in their final month (i.e., fourth or eighth) of interview and are, consequently, asked about their labor earnings. In an effort to focus on individuals of prime working age, I limit the sample to workers between the ages of 18 and 65.

I perform two sets of calculations from the MORG files. First, I compute educational attainment distributions and union membership rates for a collection of more than 200 industries, which correspond to an approximately three-digit (Standard Industrial Classification [SIC]) level of aggregation. To maximize the number of observations used for these computations, I use all individuals for whom an industry is identified and who report positive weekly earnings. Doing so produces a sample of 2,693,370 observations across the 20 years.

The second set of calculations involves a variety of earnings inequality measures based on hourly wages. Here, I further limit the sample to white males who report working at least 30 hours per week so that the sample consists entirely of workers with a strong attachment to the labor force (i.e., their primary activity is work). Doing so eliminates the need to account for the influence of race and gender on earnings and, thus, the attendant inequality. It also reduces the effects of part-time workers whose presence in the workforce from one year to the next may be heavily influenced by the business cycle. I further confine the sample to hourly wages between \$2.60 and \$150 (in year-2000 dollars) to remove any remaining outlier observations. These sample selection criteria are reasonably standard in the wage inequality literature (e.g., Katz and Murphy, 1992; Juhn, Murphy, and Pierce, 1993; Card and DiNardo, 2002). In all, 1,156,715 observations are used in the inequality calculations.

The industry coverage includes the entire private sector. As noted, industries are mostly defined at a three-digit (SIC) level of aggregation, although some two-digit sectors and combinations of either three- or four-digit sectors are also represented.[10] For example, Coal Mining (CPS

[9] If SBTC is indeed a function of the distribution of skill, the fraction of highly educated workers should capture (at least to a significant degree) the extent of SBTC within an industry.

[10] A major (one-digit) sector is also included, Construction. To provide some sense of the differences between two-, three-, and four-digit industries, Pharmaceutical Preparations and Medicinals and Botanicals are four-digit sectors that belong to the three-digit sector Drugs, which, in turn, is included in the two-digit sector Chemicals and Allied Products.

industry code 41) and Air Transportation (code 421) are two-digit sectors; Hardware Stores (code 581), Drugs (181), and Advertising (code 721) are three-digit sectors; Glass and Glass Products (code 250) is a combination of three-digit sectors; and Primary Aluminum (code 272) is a collection of four-digit sectors. A total of 230 industries are identified over the 20-year period.

To quantify computer usage rates within each industry, I use the October supplements to the CPS for the years 1984, 1989, 1993, and 1997. In these supplements, individuals were asked about their computer use at work, including (for some of the years) the types of tasks performed using this equipment. Computer usage rates are calculated as frequencies of positive responses to the question: "Do you directly use a computer at work?" Note that these are the same data used by Krueger (1993) in his study of the effect of information technology on wages and by Autor, Katz, and Krueger (1998) in their analysis of computer usage and skill distributions. Here, too, to maximize the number of observations used to estimate computer usage within detailed sectors, the calculations incorporate all workers for whom an industry of employment can be identified. Additional details about the construction of the final data set appear in the appendix.

Some Trends

As noted in the introduction, three broad features characterize the evolution of the U.S. wage distribution in recent decades: rising overall dispersion, widening gaps between workers with different levels of education, and increasing dispersion among workers with the same levels of education. All three are evident in the CPS data examined here.

In 1983, for example, the 90th percentile of the overall hourly wage distribution was roughly 1.9 times as large as the median. By 2002, it was 2.2 times as large. Figure 1 shows a similar result using an alternative measure of overall wage dispersion, the variance of log hourly wages.[11] Although there have been years in which the variance has decreased, the general trend has clearly been upward, rising 19 percent between 1983 and 2002.

Between-education-group gaps also exhibit an upward trend. High school graduates in the sample, for instance, earned roughly 16 percent more than high school dropouts in 1983.[12] In 2002, they earned 27 percent more. What is even more striking, however, is the gap at the top end of the educational attainment distribution. Figure 2 plots the evolution of the wage premium earned by college graduates relative to workers with only a high school diploma. This wage differential increased from an average of 54 percent in 1983 to 73 percent in 2002.[13]

To see that earnings differentials among workers with the *same* levels of education have also grown over this period, consider Figure 3, which plots the variance of the residuals following a regression of log hourly wages on education and experience.[14] Based on the calculations, the variance of these "residual earnings"—which is often interpreted as the degree of spread in the earnings distribution of workers with the same observable levels of skill—rose by nearly 20 percent over this 20-year period.

BETWEEN-INDUSTRY VERSUS WITHIN-INDUSTRY INEQUALITY

Overall Wages

To assess the degree to which rising U.S. earnings dispersion has been a between- or within-industry phenomenon, I consider the following straightforward decomposition. Given a sample

[11] Wages are usually expressed in logarithms in studies of labor earnings. Doing so facilitates both the computation and interpretation of the results (see Card, 1999).

[12] These figures are based on the coefficients from year-specific regressions of log hourly wages on four educational attainment dummies (no high school, some high school, high school, some college, college); a quartic polynomial in potential experience; and indicators for marital status, union membership, metropolitan status, and Census division of residence.

[13] Figure 2 plots "log point" differences (i.e., the difference between the log wages of one group and the log wages of another). To derive percentages, simply calculate ($exp(x) - 1$), where x is the log point difference.

[14] More precisely, these residuals are derived from the same regression as that described in footnote 12.

Figure 1

Overall Wage Variance of Log Hourly Wages

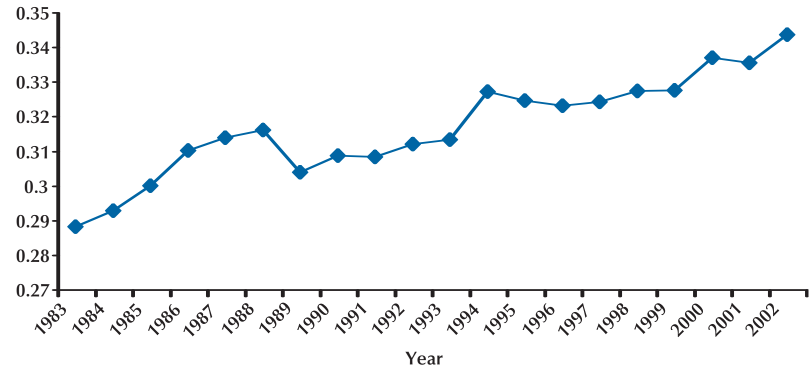


Figure 2

College Graduates Relative to High School Graduates

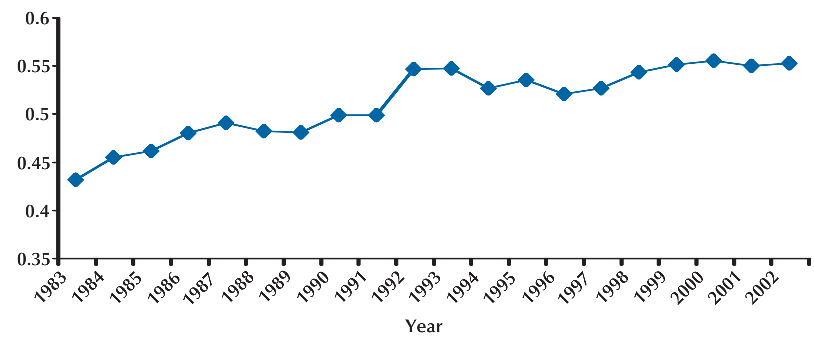


Figure 3

Residual Wage Variance of Log Hourly Wages

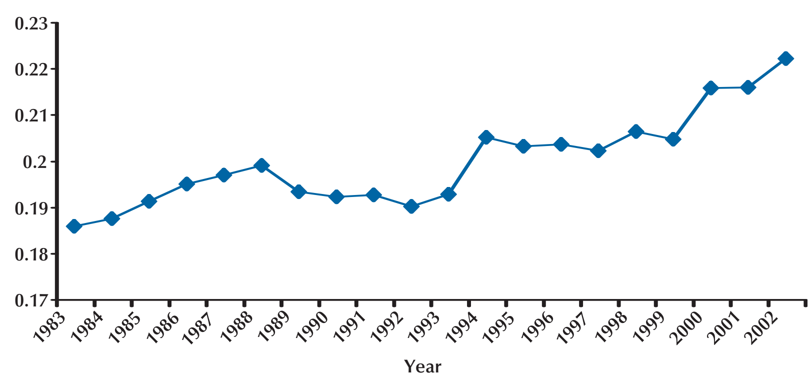


Table 1

Overall Inequality Decomposition

Year	Total wage variance	Within-industry component	Between-industry component
1983	0.288	0.222 (77.1)	0.066 (22.9)
1984	0.293	0.228 (77.9)	0.065 (22.1)
1985	0.3	0.23 (76.8)	0.069 (23.2)
1986	0.31	0.237 (76.6)	0.073 (23.4)
1987	0.313	0.24 (76.6)	0.073 (23.4)
1988	0.316	0.242 (76.8)	0.073 (23.2)
1989	0.304	0.233 (76.8)	0.07 (23.2)
1990	0.308	0.237 (76.9)	0.071 (23.1)
1991	0.308	0.235 (76.5)	0.072 (23.5)
1992	0.312	0.237 (76.1)	0.075 (23.9)
1993	0.313	0.24 (76.7)	0.073 (23.3)
1994	0.327	0.259 (79.3)	0.068 (20.7)
1995	0.324	0.259 (79.9)	0.065 (20.1)
1996	0.323	0.256 (79.2)	0.067 (20.8)
1997	0.324	0.258 (79.8)	0.066 (20.2)
1998	0.327	0.261 (79.7)	0.066 (20.3)
1999	0.327	0.262 (80)	0.065 (20)
2000	0.337	0.266 (79)	0.071 (21)
2001	0.335	0.267 (79.7)	0.068 (20.3)
2002	0.343	0.275 (80.2)	0.068 (19.8)

NOTE: Between- and within-industry components of total variance in log hourly wages as defined by equation (3). Percentages of total variance accounted for by each component are reported in parentheses. The final column is calculated by dividing the annual changes in the within-industry component by the corresponding changes in total variance.

of workers, the variance of the hourly wage distribution in year t, V_t, can be estimated as

$$(1) \qquad V_t = \frac{1}{N_t}\sum_{i=1}^{I_t}\sum_{j=1}^{N_{i,t}}(w_{j,i,t} - \bar{w}_t)^2,$$

where $w_{j,i,t}$ is the wage of worker j of industry i, $\bar{w}_t$ is the overall mean wage, $N_{i,t}$ is the number of workers in industry i, I_t is the number of industries, and N_t is the total number of workers, $\Sigma_i N_{i,t}$, all for the year t. If we rewrite this expression as

$$(2) \qquad V_t = \frac{1}{N_t}\sum_{i=1}^{I_t}\sum_{j=1}^{N_{i,t}}(w_{j,i,t} - \bar{w}_{i,t} + \bar{w}_{i,t} - \bar{w}_t)^2,$$

where $\bar{w}_{i,t}$ denotes the mean wage among workers of industry i, the variance of the wage distribution can be expressed as the sum of two terms[15]:

(3)
$$V_t = \frac{1}{N_t}\sum_{i=1}^{I_t}\sum_{j=1}^{N_{i,t}}(w_{j,i,t} - \bar{w}_{i,t})^2 + \frac{1}{N_t}\sum_{i=1}^{I_t}\sum_{j=1}^{N_{i,t}}(w_{i,t} - \bar{w}_t)^2.$$

The first, given by the sum of squared deviations of individual wages from their industry means, can be interpreted as a "within-industry" component of wage dispersion. The second, which is constructed from the sum of squared deviations of the industry means from the overall mean, can be viewed as a "between-industry" component. By calculating these two pieces, we can gain some insight into the extent to which rising wage inequality in the United States over the past two

[15] The derivation of (3) is sketched in the appendix.

Table 2

Residual Inequality Decomposition

Year	Total wage variance	Within-industry component	Between-industry component
1983	0.186	0.159 (85.7)	0.026 (14.3)
1984	0.187	0.162 (86.7)	0.025 (13.3)
1985	0.191	0.165 (86.2)	0.026 (13.8)
1986	0.195	0.167 (85.8)	0.028 (14.2)
1987	0.197	0.17 (86.4)	0.027 (13.6)
1988	0.199	0.171 (86.3)	0.027 (13.7)
1989	0.193	0.168 (87)	0.025 (13)
1990	0.192	0.168 (87.3)	0.024 (12.7)
1991	0.192	0.169 (87.7)	0.024 (12.3)
1992	0.19	0.167 (87.7)	0.023 (12.3)
1993	0.193	0.17 (88.3)	0.022 (11.7)
1994	0.205	0.184 (90)	0.021 (10)
1995	0.203	0.183 (90.2)	0.02 (9.8)
1996	0.203	0.184 (90.2)	0.02 (9.8)
1997	0.202	0.183 (90.7)	0.019 (9.3)
1998	0.206	0.187 (90.9)	0.019 (9.1)
1999	0.205	0.187 (91.5)	0.017 (8.5)
2000	0.216	0.196 (90.8)	0.02 (9.2)
2001	0.216	0.197 (91.4)	0.019 (8.6)
2002	0.222	0.204 (92)	0.018 (8)

NOTE: Between- and within-industry components of total variance in residual log hourly wages (after a regression on education and experience) as defined by equation (3). Percentages of total variance accounted for by each component are reported in parentheses. The final column is calculated by dividing the annual changes in the within-industry component by the corresponding changes in total variance.

decades has been a between- or within-industry phenomenon.

The resulting components using overall (i.e., unconditional) log hourly wages are listed in Table 1. Most obviously, they show that the vast majority of the wage dispersion observed each year is due to earnings variation within industries. This particular result can be seen from the fractions of total variation accounted for by each component, which are reported in parentheses. In all years, the fraction of total wage variance accounted for by within-industry variation is between 75 and 80 percent. More importantly, there seems to have been a gradual rise in this fraction over time. At the beginning of the sample time frame, the within-industry component averaged approximately 77 percent of the total. By 2002, it was closer to 80 percent.

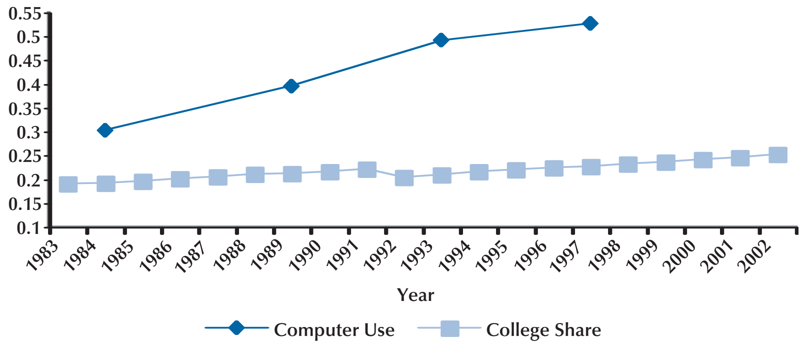
In all, the variance of log hourly wages rose from 0.29 to 0.34 between 1983 and 2002. While both the within- and between-industry components also rose over this time frame, the within-industry part accounted for nearly all—approximately 96.4 percent—of the increase in total variance.

Residual Wages

As noted previously, one of the basic features of the rise in U.S. wage inequality is the growing degree of dispersion in the wage earnings of workers with similar levels of skill (i.e., education and experience). That is, the degree of variation among,

Figure 4

Computer Usage and College Completion Rates



say, college graduates with 10 years of work experience has grown larger in recent decades. Has this rise in residual inequality also been primarily a within-industry phenomenon?

To consider this question, I begin by regressing log hourly wages on a vector of personal covariates, including years of education; four educational attainment indicators (no high school, some high school, high school, some college, college or more); interactions of years of education with these indicators; a quartic polynomial in potential work experience; and dummies for marital status, union membership, residence in a metropolitan area, and Census division.[16] I then collect the residuals and use them in place of actual wages, $w_{j,i,t}$, when calculating the within- and between-industry pieces in (3).

Table 2 shows the results. Qualitatively, they show precisely the same result as with overall wages. In each year, the within-industry component is by far the larger piece of overall variation, averaging between, roughly, 85 and 90 percent of the total. Additionally, there has been a gradual rise in this fraction over time—from 86 percent in 1983 to approximately 92 percent in 2002—suggesting that the within-industry component has become more important over time.

On the whole, between 1983 and 2002, the change in within-industry residual wage variation actually exceeded the increase in total residual wage variation. To be specific, increases in within-industry residual wage variation accounted for 123.9 percent of the change in total residual variation. Hence, there was actually a net decrease in the extent of inequality across workers possessing similar characteristics but employed in different industries over these years. Evidently, whatever has driven the rise of inequality across workers—either with similar levels of education or not—has done so primarily within individual industries.

INEQUALITY, EDUCATION, AND COMPUTER USE

Baseline Results

Given that the majority of rising wage dispersion has been the result of growing differentials among workers within the same industry, I now turn to the analysis of within-industry inequality trends. Specifically, this section examines the role of skill distributions (i.e., fractions of highly educated workers in total employment) and information technology in explaining industry-specific earnings inequality.

As shown in Figure 4, the fraction of total

16 These regressions are performed separately for each year. Again, the logarithmic transformation of wage earnings is standard in the labor literature as is the specification of the wage-experience profile by means of a fourth-order polynomial (e.g., Autor, Katz, and Krueger, 1998). In terms of years of schooling completed, the educational attainment categories correspond to 0-8 (no high school), 9-11 (some high school), 12 (high school), 13-15 (some college), 16+ (college).

Table 3

Education and Computer Usage Changes by Major Sector

Industry	College employment fraction change 1983-2002	Computer usage change 1984-97
Agriculture, forestry, fisheries	0.02	0.17
Mining	–0.06	0.07
Construction	–0.001	0.13
Nondurable manufacturing	0.09	0.23
Durable manufacturing	0.05	0.15
Transportation	0.055	0.19
Wholesale trade	0.04	0.18
Retail trade	0.013	0.21
Finance, insurance, real estate	0.11	0.24
Business and repair services	0.11	0.29
Personal services	0.04	0.19
Entertainment and recreation services	0.05	0.26
Professional and related services	0.05	0.32

NOTE: Changes in proportions of employees with a bachelor's degree and using a computer at work.

employment accounted for by college graduates and the fraction of workers who use a computer at work have both increased in the past two decades. In 1983, the fraction of workers with at least a bachelor's degree stood at 19 percent. By 2002, it had grown to 25 percent. Similarly, between 1984 and 1997, the fraction of workers reporting use of a computer at work increased from 30 percent to approximately 53 percent.

Table 3 shows that both of these qualitative patterns were reasonably widespread, at least in the sense that they occurred in nearly every major industrial sector. In fact, of the 13 sectors listed in the table, only 2—Mining and Construction—saw their college employment fractions decrease over the sample period. All, as it happens, witnessed increasing computer usage. Among the 228 (of 230) more detailed industry groups identified in both the beginning and ending years of the sample, the results are similar. A total of 175 increased their fraction of college graduates in total employment between 1983 and 2002, while 225 saw increases in their computer usage rates between 1984 and 1997.[17]

To what extent do these trends account for industry-specific levels of wage inequality? I attempt to draw inferences about the answer using the following statistical characterization of inequality in industry i in year t, $Ineq_{i,t}$:

$$Ineq_{i,t} = \alpha + \delta_t + \beta X_{i,t} + \varepsilon_{i,t}, \tag{4}$$

where α is an overall constant; δ_t represents a time dummy added to capture the temporal variation in inequality evident from Figures 1, 2, and 3[18]; $X_{i,t}$ is a vector of time-varying industry characteristics; and $\varepsilon_{i,t}$ is a residual. Three quantities are included in $X_{i,t}$: the fraction of workers with a bachelor's degree,[19] the fraction using a computer

[17] The mean (standard deviation) change in the college employment fraction is 0.04 (0.08); the change in the computer usage rate is 0.25 (0.16).

[18] I reestimated specification I of equation (4) further adding industry-specific time trends to capture differences in the temporal behavior of inequality across industries. The resulting estimates did not differ substantially from those reported here. All of the college-share coefficients were significantly positive; all but one of the union rate coefficients were significantly negative.

[19] I also considered the share of total work hours accounted for by college-educated workers instead of the college employment fraction. Since the correlation between these two variables exceeds 0.99, the results did not differ substantially from what is reported here.

at work, and the fraction who are members of a union. As suggested previously, computer usage is intended as a direct measure of SBTC, whereas the college share is used as an indirect measure. Both interpretations seem justifiable in light of the research surveyed here connecting technological adoption and the distribution of education/skill. The unionization rate is included to capture the influence of an institutional characteristic that has likely contributed to changes in industry-level inequality. Since these independent variables are calculated from the CPS micro samples, I restrict estimation to those industry-years for which at least 10 observations were available, in an effort to reduce the sampling noise inherent in each.

Because the computer usage data are available for only 4 of the 20 years, I consider three different specifications of this equation. In the first, I limit $X_{i,t}$ to the college and unionization fractions so that I am able to use all 20 years of data. Direct evidence correlating inequality with technology is then given in the second specification, which drops the college fraction but adds the computer usage rate. The third specification considers both direct and indirect measures of SBTC simultaneously by adding the college employment fraction to this second specification.

To keep the analysis as broad as possible, I examine three different categories of inequality measures: (i) overall, (ii) between-education-group, and (iii) residual. The overall measures include the variance of log hourly wages and the differences between the 90th, 50th, and 10th percentiles of the log hourly wage distribution. The between-education-group gaps are given by differences between the average log wages of college graduates and those in each of the following four categories: some college, high school, some high school, no high school. Residual inequality is given by the same statistics considered for overall inequality, where the calculations are done using the residuals following the regressions described in the subsection "Residual Wages."

In all cases, estimation proceeds by generalized least squares in which the industry-year observations are weighted by the number of CPS observations used to calculate the inequality variables. An inequality figure based on 10 observations, after all, ought to involve greater sampling error than one based on 1,000. This weighting procedure helps to account for the differential degree of noise across observations.

Results appear in Table 4. On the whole, they show that inequality tends to be significantly associated with each of the three regressors—the college employment fraction, the computer usage rate, and the extent of union membership. The unionization rate, as expected, enters negatively in 34 (of 36) instances (significantly in 30), suggesting that decreasing unionization is an important piece of the rise in nearly every wage gap considered. Such a result, of course, reinforces the general view established in the inequality literature that the decline in union activity in the United States has been a major element in the rise of earnings disparity.[20]

The two measures of SBTC, by contrast, both enter positively in nearly every case. Indeed, none of the college fraction coefficients and only one of the estimated computer usage coefficients are negative (albeit statistically insignificant). What is more, of the 24 coefficients for each of these two quantities, a large number are statistically important: 20 of the college fraction coefficients and 16 of the computer usage coefficients. The majority of the insignificant coefficients, incidentally, emerge from specification III in which both of these regressors appear. Quite possibly, the lack of significance in these cases derives from the strong correlation between these two variables.[21]

Are the estimated associations *economically* important? Just focusing on the overall 90-10 wage difference, the point estimates suggest that a 1-standard-deviation increase in either the college employment fraction or the computer usage rate is accompanied by a 4- to 7-percentage-point increase in this differential.[22] When evaluated at

[20] Blau and Kahn (1996), for instance, find that differences in labor market institutions (e.g., unionization) account for a large part of the difference between inequality in the United States and that of nine other OECD countries.

[21] The correlation between the college employment fraction and the frequency of computer usage is 0.61, which is consistent with the findings of Autor, Katz, and Krueger (1998).

[22] Summary statistics for the variables used in the inequality regressions are reported in Table A1 of the appendix.

Table 4

Baseline Inequality Regressions

Dependent variable	Specification	College fraction	Computer rate	Union rate	R^2
Overall variance	I	0.18* (0.01)	—	–0.17* (0.008)	0.38
	II	—	0.11* (0.01)	–0.23* (0.02)	0.27
	III	0.1* (0.03)	0.06* (0.02)	–0.22* (0.02)	0.31
Overall 90-10 difference	I	0.53* (0.03)	—	–0.52* (0.03)	0.35
	II	—	0.31* (0.05)	–0.66* (0.07)	0.24
	III	0.31* (0.1)	0.17* (0.07)	–0.63* (0.07)	0.29
Overall 90-50 difference	I	0.1* (0.02)	—	–0.55* (0.02)	0.33
	II	—	0.002 (0.03)	–0.62* (0.04)	0.25
	III	0.02 (0.06)	–0.006 (0.04)	–0.64* (0.05)	0.26
Overall 50-10 difference	I	0.43* (0.02)	—	0.03 (0.02)	0.26
	II	—	0.31* (0.04)	–0.04 (0.06)	0.2
	III	0.29* (0.05)	0.18* (0.04)	0.005 (0.06)	0.28
College–no high school	I	0.58* (0.05)	—	–0.002 (0.04)	0.19
	II	—	0.42* (0.06)	–0.08 (0.07)	0.19
	III	0.15 (0.16)	0.35* (0.1)	–0.06 (0.08)	0.2
College–some high school	I	0.5* (0.04)	—	–0.14* (0.02)	0.21
	II	—	0.3* (0.04)	–0.16* (0.04)	0.2
	III	0.23* (0.13)	0.2* (0.08)	–0.13* (0.04)	0.2
College–high school	I	0.17* (0.02)	—	–0.18* (0.02)	0.09
	II	—	0.12* (0.04)	–0.2* (0.05)	0.09
	III	0.13 (0.09)	0.06 (0.06)	–0.18* (0.05)	0.09
College–some college	I	0.09* (0.02)	—	–0.11* (0.02)	0.08
	II	—	0.054* (0.027)	–0.14* (0.03)	0.07
	III	0.07 (0.07)	0.02 (0.04)	–0.13* (0.03)	0.07
Residual variance	I	0.11* (0.008)	—	–0.12* (0.006)	0.34
	II	—	0.05* (0.01)	–0.17* (0.02)	0.21
	III	0.08* (0.03)	0.01 (0.02)	–0.16* (0.02)	0.26
Residual 90-10 difference	I	0.32* (0.02)	—	–0.41* (0.02)	0.32
	II	—	0.16* (0.04)	–0.52* (0.06)	0.18
	III	0.23* (0.07)	0.05 (0.05)	–0.5* (0.06)	0.23
Residual 90-50 difference	I	0.16* (0.02)	—	–0.26* (0.01)	0.26
	II	—	0.08* (0.02)	–0.32* (0.04)	0.15
	III	0.12* (0.05)	0.03 (0.03)	–0.32* (0.04)	0.19
Residual 50-10 difference	I	0.15* (0.01)	—	–0.15* (0.01)	0.21
	II	—	0.07* (0.02)	–0.2* (0.03)	0.11
	III	0.11* (0.04)	0.03 (0.03)	–0.19* (0.03)	0.14

NOTE: Coefficients from estimation of (4). Specification I uses annual data 1983-2002. Specifications II and III use only data from 1984, 1989, 1993, and 1997. Each regression includes year dummies. Heteroskedasticity-consistent standard errors are reported in parentheses; * denotes significance at the 10 percent level.

the mean value in dollar terms, this figure represents an increase of approximately $0.75 to $1.31 in the 90-10 differential. Such a magnitude is roughly one-quarter of the standard deviation of the 90-10 differentials in the sample.

Looking at the between-education-group gaps, there are some sizable associations here too. A 1-standard-deviation increase in the college fraction, for example, correlates with a 3- to 7-percentage-point ($0.27 to $0.62) increase in the college–some high school gap and a 2- to 2.5-percentage-point ($0.14 to $0.18) increase in the college–high school gap. For computer usage, a 1-standard-deviation increase is associated with a 4.5- to 7-percentage-point ($0.40 to $0.62) rise in the college–some high school difference and a 1.5 to 2.5 percentage point ($0.11 to $0.18) increase in the college–high school gap. Given sample standard deviations of, respectively, 36 and 30 percentage points for the college–some high school and college–high school gaps, these correlations are far from trivial.

There is, to be sure, some variation in the coefficient magnitudes and statistical significance of these two variables across the inequality measures. In particular, they tend to have the largest associations with those measures involving the position of the bottom of the distribution relative to the middle and top, at least when considering the overall and between-education-group gaps. For example, for the overall percentile differences, the coefficients are much larger for the 50-10 differentials than they are for the 90-50 differentials. The positive associations between the college and computer usage fractions and the overall 90-10 differential, therefore, clearly seem to be driven by the bottom half of the wage distribution.

A similar result can be inferred from the between-education-group measures, which show larger coefficients on the two SBTC variables when considering the two "top-bottom" inequality variables (college–no high school, college–some high school) than when looking at the two "top-middle" variables (college–high school, college–some college). This pattern is consistent with the idea that workers at the bottom ends of the wage and educational attainment distributions were the hardest hit by new technologies (or, at least, benefited the least from them).

Endogeneity Considerations

Although the regressors have been treated as exogenous thus far, it is possible that one in particular may be endogenous with respect to inequality: education. The fraction of an industry's workers with a college degree may, for example, be an increasing function of the relative returns paid to these workers. Hence, a rise in the college–high school gap could increase the college fraction—either by attracting more college graduates or driving away high school graduates, depending on what causes the gap to increase—which would bias the estimated coefficient on the college fraction upward.

In an effort to address this possibility, I consider the following simple exercise. I regress the annual changes in an industry's college employment fraction on the initial levels of inequality and a set of year dummies.[23] I then make inferences about the extent to which inequality influences the college fraction by examining the coefficients on inequality. A significant coefficient, naturally, would suggest that inequality levels have a non-negligible influence on the educational mix of workers.

The first column of figures in Table 5 shows the results, which have a nearly uniform lack of significance of initial inequality in explaining subsequent changes in industry-specific college fractions: Only 2 of the 12 coefficients are significant. These results seem to cast some doubt on the notion that an industry's college employment fraction is endogenous with respect to inequality.

Of course, because this specification may not adequately capture the response of education to changes in inequality, I also consider an alternative in which changes in the college employment fraction are regressed on one lag of the change in inequality (i.e., the change of an industry's college fraction between 2000 and 2001 is regressed on the change in its level of inequality between 1999 and 2000). Hence, instead of correlating subsequent changes in education with initial *levels* of inequality, this equation estimates how changes in education are associated with recent *changes*

[23] As before, I restrict these regressions to industry-year observations based on at least 10 observations for the inequality calculations.

Table 5

Education as a Function of Inequality

Measure	Specification: Initial levels	Specification: Lagged differences
Overall variance	–0.0007 (0.009)	–0.02* (0.01)
Overall 90-10 difference	0.0002 (0.0003)	–0.008* (0.005)
Overall 90-50 difference	–0.0003 (0.005)	–0.007* (0.005)
Overall 50-10 difference	0.0008 (0.005)	–0.005 (0.007)
College–no high school	–0.002 (0.002)	–0.006* (0.002)
College–some high school	0.007* (0.003)	0.001 (0.002)
College–high school	0.005 (0.004)	0.001 (0.002)
College–some college	0.005 (0.005)	–0.002 (0.004)
Residual variance	0.005 (0.01)	0.005 (0.02)
Residual 90-10 difference	0.003 (0.005)	0.003 (0.005)
Residual 90-50 difference	0.01* (0.006)	0.003 (0.006)
Residual 50-10 difference	–0.004 (0.01)	0.003 (0.008)

NOTE: Coefficients on inequality, in both initial levels and lagged first differences, from regressions of the annual change in an industry's college fraction on inequality. Regressions also include year dummies. Heteroskedasticity-consistent standard errors are reported in parentheses; * denotes significance at the 10 percent level.

in inequality. Those results appear in the second column of figures in Table 5.

Here, interestingly, a greater number of the coefficients—4 of the 12—are significantly nonzero at conventional levels (i.e., at least 10 percent). However, of these, all are negative, indicating that increases in an industry's inequality tend to be followed by decreases in its college employment fraction. This particular result implies that, if anything, the coefficients listed in Table 4 may actually be biased downward (i.e., toward zero) and, thus, understate the association between education and inequality. Although certainly not definitive, I take this evidence as *suggesting* that endogenous education does not pose a significant problem for the qualitative interpretation of the results.

Alternative Specifications

This section considers two alterations of the analysis described here. In the first, I look at the possibility that the dispersion of computer usage, rather than the mean usage rate, influences the degree of wage inequality. Indeed, there could very well be a nonlinear relationship between the computer usage rate and the degree of spread in the wage distribution. It may be, for instance, that the relationship is positive at low levels of computer usage, but negative as the usage rate closes in on unity.

The second augments the regression considered in (4) with industry-specific fixed effects:

$$Ineq_{i,t} = \alpha + \delta_t + \mu_i + \beta X_{i,t} + \varepsilon_{i,t}, \tag{5}$$

where μ_i is a constant element influencing the degree of earnings inequality in industry *i*. Doing so controls for all time-invariant industry characteristics, unobserved or otherwise, that may influence inequality and, thus, eliminates any bias resulting from the omission of these characteristics.[24]

[24] One could also treat the industry-specific terms as stochastic and estimate (5) by random effects. However, consistency of the random effects estimator depends on the assumption that these terms are uncorrelated with the regressors (see Wooldridge, 2002, p. 257). Because the fixed-effects estimator is consistent whether this condition is satisfied or not, I treat the μ_i as a set of constants to be estimated.

Table 6

Inequality and the Variance of Computer Usage

Dependent variable	College fraction	Computer rate	Computer variance
Overall variance	0.08* (0.03)	0.05* (0.02)	0.2* (0.06)
Overall 90-10 difference	0.26* (0.1)	0.13* (0.07)	0.61* (0.18)
Overall 90-50 difference	–0.01 (0.06)	–0.03 (0.04)	0.37* (0.11)
Overall 50-10 difference	0.27* (0.06)	0.16* (0.04)	0.24* (0.12)
College–no high school	0.16 (0.16)	0.32* (0.12)	0.1 (0.3)
College–some high school	0.24* (0.13)	0.16* (0.08)	0.14 (0.16)
College–high school	0.13 (0.09)	0.01 (0.05)	0.48* (0.12)
College–some college	0.07 (0.07)	0.01 (0.04)	0.29* (0.09)
Residual variance	0.075* (0.03)	–0.0004 (0.02)	0.11* (0.04)
Residual 90-10 difference	0.2* (0.08)	0.03 (0.05)	0.32* (0.13)
Residual 90-50 difference	0.11* (0.05)	0.01 (0.04)	0.2* (0.09)
Residual 50-10 difference	0.1* (0.04)	0.02 (0.03)	0.12* (0.07)

NOTE: Coefficients from estimation of Specification III of (4) in which the variance of computer usage has also been added. Heteroskedasticity-consistent standard errors are reported in parentheses; * denotes significance at the 10 percent level.

Table 6 shows results from the inclusion of the variance of computer usage. For the sake of brevity, I have reported only the output from specification III, in which $X_{i,t}$ contains the college graduate, computer usage, and union membership fractions. While there is a slight dropoff in some of the magnitudes of the coefficients relative to the baseline estimates in Table 4, most are little changed after including the variance of computer usage. In fact, the same coefficients that are significant in Table 4 are significant here as well.

Computer use variance itself also enters positively and, for the most part, significantly, just as one would expect. Each of the 12 coefficients reported in the table is positive, 10 significantly so. In these data, then, both the first and second moments of the distribution of computer usage correlate directly with wage inequality.[25]

Table 7 reports the results when both time- and industry-specific fixed effects are included in the regressions. Two features of the results are especially notable. First, the majority of the coefficients on both the college employment fraction and frequency of computer use are positive, while those on the unionization rate are negative, just as in the baseline results. However, second, the number of coefficients that differ statistically from zero at conventional levels has dropped relative to the results in Table 4. To be sure, among the college fraction coefficients, more than half (15 of 24) remain significant, indicating that industry-specific changes in many of the inequality measures correlate strongly with changes in their fractions of highly educated labor. At the same time, only 13 of the 36 unionization coefficients and 3 of the 24 computer usage coefficients differ significantly from zero after conditioning on time-invariant industry terms.

Very likely, this decrease in significance stems from the decline in the extent of variation in the data once industry-specific intercepts are included. This particular aspect of the estimation can be inferred from the sharp rise in the goodness-of-fit statistics reported in the final columns of Tables 4 and 7. The average R^2 rises from 0.21 to 0.67 with the addition of the industry effects.

[25] Since the majority of industry-year observations have computer usage fractions less than 0.5, it is not surprising that the mean and variance of the computer usage distribution are positively associated. However, the correlation is relatively modest, 0.47, suggesting that each variable may reasonably pick up some of the variation in inequality independently of the other.

Table 7

Inequality Regressions—Industry Effects Included

Dependent variable	Specification	College fraction	Computer rate	Union rate	R^2
Overall variance	I	0.25* (0.02)	—	–0.09* (0.02)	0.8
	II	—	0.08* (0.05)	–0.16* (0.09)	0.8
	III	0.24* (0.08)	0.07 (0.05)	–0.12 (0.09)	0.81
Overall 90-10 difference	I	0.68* (0.12)	—	–0.5* (0.09)	0.77
	II	—	0.08 (0.17)	–0.23 (0.4)	0.76
	III	1.07* (0.4)	0.03 (0.17)	–0.13 (0.4)	0.77
Overall 90-50 difference	I	0.17* (0.1)	—	–0.34* (0.07)	0.67
	II	—	–0.04 (0.11)	–0.03 (0.25)	0.7
	III	0.28 (0.4)	–0.06 (0.12)	0.02 (0.3)	0.7
Overall 50-10 difference	I	0.5* (0.06)	—	–0.16* (0.06)	0.69
	II	—	0.12 (0.12)	–0.2 (0.24)	0.72
	III	0.79* (0.16)	0.09 (0.12)	–0.16 (0.22)	0.74
College–no high school	I	0.08 (0.14)	—	–0.24* (0.11)	0.59
	II	—	0.06 (0.15)	–0.18 (0.27)	0.67
	III	0.32 (0.4)	0.03 (0.16)	–0.13 (0.3)	0.67
College–some high school	I	0.17 (0.11)	—	–0.18* (0.08)	0.6
	II	—	0.25* (0.12)	–0.32* (0.19)	0.67
	III	0.27 (0.3)	0.25* (0.12)	–0.29 (0.2)	0.67
College–high school	I	0.04 (0.07)	—	–0.16* (0.06)	0.6
	II	—	0.1 (0.08)	–0.15 (0.13)	0.7
	III	0.1 (0.2)	0.1 (0.08)	–0.12 (0.14)	0.7
College–some college	I	0.007 (0.07)	—	–0.06 (0.06)	0.44
	II	—	0.095 (0.08)	0.05 (0.14)	0.57
	III	–0.17 (0.16)	0.11 (0.08)	0.007 (0.14)	0.57
Residual variance	I	0.13* (0.02)	—	–0.04* (0.02)	0.74
	II	—	0.04 (0.03)	–0.13 (0.08)	0.7
	III	0.13* (0.07)	0.03 (0.03)	–0.16* (0.08)	0.7
Residual 90-10 difference	I	0.44* (0.1)	—	–0.16* (0.07)	0.7
	II	—	0.03 (0.14)	0.28 (0.3)	0.69
	III	0.78* (0.3)	–0.01 (0.1)	0.17 (0.3)	0.7
Residual 90-50 difference	I	0.21* (0.07)	—	–0.03 (0.05)	0.59
	II	—	–0.08 (0.11)	0.09 (0.22)	0.58
	III	0.43* (0.2)	–0.1 (0.1)	–0.02 (0.2)	0.59
Residual 50-10 difference	I	0.23* (0.06)	—	–0.13* (0.05)	0.53
	II	—	0.11 (0.07)	0.2 (0.17)	0.59
	III	0.35* (0.18)	0.09 (0.07)	0.19 (0.16)	0.59

NOTE: Coefficients from estimation of (5). Specification I uses annual data 1983-2002. Specifications II and III use only data from 1984, 1989, 1993, and 1997. Each regression includes year dummies and industry-specific fixed effects. Heteroskedasticity-consistent standard errors are reported in parentheses; * denotes significance at the 10 percent level.

While this necessarily tempers the conclusions that can be drawn from the results, one can still interpret this evidence as consistent with the SBTC hypothesis.

CONCLUSIONS

Despite the presence of a large literature examining the rise of earnings inequality in the United States, surprisingly few studies have directly explored the role of information technology in driving this trend. Such an omission is particularly surprising in light of the general consensus that has emerged in support of the skill-biased technological change hypothesis. This paper has attempted to offer some evidence on this issue.

The findings indicate that the vast majority of the rise in U.S. wage inequality over the past two decades is the product of increasing gaps between workers within the same industry rather than between workers across different industries. This result holds whether considering workers of differing levels of observable skill (overall inequality) or those with the same levels (residual inequality). What is more, within-industry inequality—defined in overall, residual, and between-education-group terms—tends to be positively associated with the two measures of skill-biased technological change considered here, the college employment fraction and the frequency of computer usage. Collectively, these two observations are compatible with the idea that skill-biased technological change has been a significant element in the rise of wage dispersion in the United States.

Of course, because the two measures of skill-biased technological change considered here are less than ideal, there remains ample room for additional research on this topic. In particular, studies examining the extent to which plants and industries have adopted specific production technologies, such as those considered by Dunne (1994), and how the implementation of those technologies correlate with earnings differentials would greatly assist in clarifying the skill-biased technological change hypothesis. Considering the popularity of the theory, such an undertaking certainly seems worthwhile.

REFERENCES

Acemoglu, Daron. "Changes in Unemployment and Wage Inequality: An Alternative Theory and Some Evidence." *American Economic Review*, December 1999, *89*(5), pp. 1259-78.

Acemoglu, Daron. "Technical Change, Inequality, and the Labor Market." *Journal of Economic Literature*, March 2002, *40*(1), pp. 7-72.

Autor, David H.; Katz, Lawrence F. and Krueger, Alan B. "Computing Inequality: Have Computers Changed the Labor Market?" *Quarterly Journal of Economics*, November 1998, *113*(4), pp. 1169-213.

Autor, David H.; Levy, Frank and Murnane, Richard J. "The Skill Content of Recent Technological Change: An Empirical Exploration." *Quarterly Journal of Economics*, November 2003, *118*(4), pp. 1279-333.

Berman, Eli; Bound, John and Griliches, Zvi. "Changes in the Demand for Skilled Labor within U.S. Manufacturing: Evidence from the Annual Survey of Manufactures." *Quarterly Journal of Economics*, May 1994, *109*(2), pp. 367-97.

Blau, Francine D. and Kahn, Lawrence M. "International Differences in Male Wage Inequality: Institutions versus Market Forces." *Journal of Political Economy*, August 1996, *104*(4), pp. 791-837.

Borjas, George J.; Freeman, Richard B. and Katz, Lawrence F. "How Much Do Immigration and Trade Affect Labor Market Outcomes?" *Brookings Papers on Economic Activity*, 1997, (1), pp. 1-90.

Card, David. "The Causal Effect of Education on Earnings," in Orley Ashenfelter and David Card, eds., *Handbook of Labor Economics*. Volume 3A. Amsterdam: Elsevier, 1999, pp. 1801-63.

Card, David and DiNardo, John E. "Skill-Biased Technological Change and Rising Wage Inequality: Some Problems and Puzzles." *Journal of Labor Economics*, October 2002, *20*(4), pp. 733-83.

Caselli, Francesco. "Technological Revolutions." *American Economic Review*, March 1999, *89*(1), pp. 78-102.

Doms, Mark; Dunne, Timothy and Troske, Kenneth R. "Workers, Wages, and Technology." *Quarterly Journal of Economics*, February 1997, *112*(1), pp. 253-90.

Dunne, Timothy. "Plant Age and Technology Use in U.S. Manufacturing Industries." *RAND Journal of Economics*, Autumn 1994, *25*(3), pp. 488-99.

Dunne, Timothy; Foster, Lucia; Haltiwanger, John and Troske, Kenneth R. "Wage and Productivity Dispersion in United States Manufacturing: The Role of Computer Investment." *Journal of Labor Economics*, April 2004, *22*(2), pp. 397-429.

Fortin, Nicole M. and Lemieux, Thomas. "Institutional Changes and Rising Wage Inequality: Is There a Linkage?" *Journal of Economic Perspectives*, Spring 1997, *11*(2), pp. 75-96.

Johnson, George E. "Changes in Earnings Inequality: The Role of Demand Shifts." *Journal of Economic Perspectives*, Spring 1997, *11*(2), pp. 41-54.

Juhn, Chinhui; Murphy, Kevin M. and Pierce, Brooks. "Wage Inequality and the Rise in Returns to Skill." *Journal of Political Economy*, June 1993, *101*(3), pp. 410-42.

Katz, Lawrence F. and Murphy, Kevin M. "Changes in Relative Wages, 1963-1987: Supply and Demand Factors." *Quarterly Journal of Economics*, February 1992, *107*(1), pp. 35-78.

Krueger, Alan B. "How Computers Have Changed the Wage Structure: Evidence from Microdata, 1984-1989." *Quarterly Journal of Economics*, February 1993, *108*(1), pp. 33-60.

Levy, Frank and Murnane, Richard J. "U.S. Earnings Levels and Earnings Inequality: A Review of Recent Trends and Proposed Explanations." *Journal of Economic Literature*, September 1992, *30*(3), pp. 1333-81.

Park, Jim Heum. "Estimation of Sheepskin Effects and Returns to Schooling Using the Old and the New CPS Measures of Educational Attainment." Working Paper No. 338, Industrial Relations Section, Princeton University, December 1994.

Topel, Robert H. "Factor Proportions and Relative Wages: The Supply-Side Determinants of Wage Inequality." *Journal of Economic Perspectives*, Spring 1997, *11*(2), pp. 55-74.

Wheeler, Christopher H. "Cities, Skills, and Inequality." Working Paper No. 2004-020A, Federal Reserve Bank of St. Louis, September 2004.

Wooldridge, Jeffrey M. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press, 2002.

APPENDIX

Data

Hourly wages are calculated as the ratio of a worker's weekly earnings to usual hours worked per week. The MORG files do topcode weekly earnings at various points ($999 for 1983 to 1988, $1,923 for 1989 to 1997, and $2,884 for 1998 to 2002). For topcoded values, I follow Card and DiNardo (2002) and impute the weekly wages as 1.4 times the topcode value to approximate the mean of the upper tail of the wage distribution. Similar techniques have been used by Katz and Murphy (1992), Juhn, Murphy, and Pierce (1993), and Autor, Katz, and Krueger (1998). All hourly wages are converted to real terms ($2,000) using the personal consumption expenditure chain-type price index. Once these values are computed, I then restrict the sample to workers with hourly wage earnings between $2.60 (which is slightly in excess of one-half the current federal minimum wage) and $150 to eliminate outliers. Inequality calculations are computed using the CPS "earnings" weight. For the educational attainment and union membership rates, a total of 2,693,370 observations are used. This corresponds to an average of 604.1 observations per industry-year (minimum = 2, maximum = 9,672). For the inequality calculations, a total of 1,156,715 observations are used. This corresponds to 258.8 observations per industry-year (minimum = 2, maximum = 7,090).

Since educational attainment is not coded in the CPS as years of schooling completed for all years (i.e., there was a change in the education variable between 1991 and 1992), I follow previous work (e.g., Autor, Katz, and Krueger, 1998) and impute values from Table 5 of Park (1994), which establishes a correspondence between the old and new CPS education variables. Potential work experience is then calculated as the maximum of 0 and (age-years of education – 6).

Computer usage by industry is calculated using responses to the question: "Do you directly use a computer at work?" reported in the October supplements for the years 1984, 1989, 1993, and 1997. Here, I use the CPS "supplement" weights for 1984, 1989, and 1993 and the "final" weight for 1997. In all, there are 59,642 observations for 1984, 60,304 for 1989, 54,273 for 1993, and 50,478 for 1997. The mean number of observations per industry-year is 308.2 (minimum = 7, maximum = 4,298).

A consistent set of 230 industries (ranging in number from 219 to 230 in any given year) are identified over the 20-year period. Because the CPS industry codes changed in 1992, a consistent classification scheme was implemented using the crosswalk provided by the U.S. Bureau of the Census (and summarized by Barry Hirsch at his website: www.trinity.edu/bhirsch).

Derivation of Variance Expression

To show that equations (1) and (3) give equivalent expressions for the variance of wages, note first that (1) expands to

$$(1') \qquad V_t = \frac{1}{N_t}\sum_{i=1}^{I_t}\sum_{j=1}^{N_{i,t}} w_{j,i,t}^2 - \frac{2}{N_t}\sum_{i=1}^{I_t}\sum_{j=1}^{N_{i,t}} w_{j,i,t}\bar{w}_t + \frac{1}{N_t}\sum_{i=1}^{I_t}\sum_{j=1}^{N_{i,t}} \bar{w}_t^2,$$

whereas (3) can be written as

$$(3') \qquad V_t = \frac{1}{N_t}\sum_{i=1}^{I_t}\sum_{j=1}^{N_{i,t}} w_{j,i,t}^2 - \frac{2}{N_t}\sum_{i=1}^{I_t}\sum_{j=1}^{N_{i,t}} w_{j,i,t}\bar{w}_{i,t} + \frac{2}{N_t}\sum_{i=1}^{I_t}\sum_{j=1}^{N_{i,t}} \bar{w}_{i,t}^2 - \frac{2}{N_t}\sum_{i=1}^{I_t}\sum_{j=1}^{N_{i,t}} \bar{w}_t\bar{w}_{i,t} + \frac{1}{N_t}\sum_{i=1}^{I_t}\sum_{j=1}^{N_{i,t}} \bar{w}_t^2 .$$

Because

$$\frac{2}{N_t}\sum_{i=1}^{I_t}\sum_{j=1}^{N_{i,t}} w_{j,i,t}\bar{w}_{i,t} = \frac{2}{N_t}\sum_{i=1}^{I_t} \bar{w}_{i,t}\bar{w}_{i,t}N_{i,t} = \frac{2}{N_t}\sum_{i=1}^{I_t} \bar{w}_{i,t}^2 N_{i,t}$$

and

$$\frac{2}{N_t}\sum_{i=1}^{I_t}\sum_{j=1}^{N_{i,t}} \bar{w}_{i,t}^2 = \frac{2}{N_t}\sum_{i=1}^{I_t} \bar{w}_{i,t}^2 N_{i,t},$$

the second and third terms on the right-hand side of (3′) sum to zero, leaving

$$(3'') \qquad V_t = \frac{1}{N_t}\sum_{i=1}^{I_t}\sum_{j=1}^{N_{i,t}} w_{j,i,t}^2 - \frac{2}{N_t}\sum_{i=1}^{I_t}\sum_{j=1}^{N_{i,t}} \bar{w}_t\bar{w}_{i,t} + \frac{1}{N_t}\sum_{i=1}^{I_t}\sum_{j=1}^{N_{i,t}} \bar{w}_t^2 .$$

Given that the middle term on the right-hand side of (3″) can be expressed as

$$\frac{2w_t}{N_t}\sum_{i=1}^{I_t} w_{i,t}N_{i,t} = \frac{2w_t}{N_t}\sum_{i=1}^{I_t} N_{i,t}\frac{1}{N_{i,t}}\sum_{j=1}^{N_{i,t}} w_{j,i,t} = \frac{2w_t}{N_t}\sum_{i=1}^{I_t}\sum_{j=1}^{N_{i,t}} w_{j,i,t} = \frac{2}{N_t}\sum_{i=1}^{I_t}\sum_{j=1}^{N_{i,t}} w_{j,i,t}w_t,$$

equations (1′) and (3′) are equivalent. Therefore, (1) and (3) are equivalent.

Table A1

Summary Statistics

Variable	Mean	Standard deviation	Minimum	Maximum	Observations
Overall variance	0.25	0.09	0	0.97	4,381
Overall 90-10 difference	1.26	0.27	0	2.82	4,381
Overall 90-50 difference	0.67	0.2	0	2.64	4,381
Overall 50-10 difference	0.59	0.17	0	1.76	4,381
College–no high school	0.68	0.35	–1.36	2.05	3,521
College–some high school	0.61	0.31	–1.17	2.02	3,943
College–high school	0.45	0.26	–0.96	2.14	4,251
College–some college	0.34	0.25	–0.81	2.51	4,263
Residual variance	0.18	0.07	0	0.67	4,379
Residual 90-10 difference	1.03	0.22	0	2.53	4,379
Residual 90-50 difference	0.52	0.15	0	1.82	4,379
Residual 50-10 difference	0.51	0.14	0	1.86	4,379
College fraction	0.21	0.14	0	0.75	4,384
Computer usage rate	0.38	0.23	0	1	887
Union membership rate	0.13	0.13	0	0.84	4,384

NOTE: Summary statistics for selected industry characteristics over the period 1983-2002.

Monetary Policy and Commodity Futures

Michelle T. Armesto and William T. Gavin

This paper constructs daily measures of the real interest rate and expected inflation using commodity futures prices and the term structure of Treasury yields. We find that commodity futures markets respond to surprise increases in the federal funds rate target by raising the inflation rate expected over the next 3 to 9 months. There is no evidence that the real interest rate responds to surprises in the federal funds target. The data from the commodity futures markets are highly volatile; we show that one can substantially reduce the noise using limited information estimators such as the median change. Nevertheless, the basket of commodities actually traded daily is quite narrow and we do not know whether our observable rates are closely connected to the unobservable inflation and real rates that affect economywide consumption and investment decisions.

Federal Reserve Bank of St. Louis *Review*, May/June 2005, *87*(3), pp. 395-405.

The Federal Reserve targets the interest rate on federal funds to implement monetary policy. The interest rate is composed of two unobservable factors, the real interest rate and a premium for expected inflation, which are important for understanding the appropriate setting of the target. Knowing how these two factors change in response to changes in the target is also important for implementing monetary policy. Empirical evidence about the level and changes in these factors is complicated by the lack of direct observations on them.[1] In this paper, we extract measures of the interest rate and expected inflation from commodity futures prices and use these measures to examine how interest rates and expected inflation respond to monetary policy shocks. Throughout this paper we use the terms inflation and real interest rate interchangeably with commodity price inflation and commodity own rate. Whether our results have important implications for monetary policy depends on how closely our measures derived from commodity markets are connected to the inflation rates and real interest rates that matter for consumption and investment decisions.

Since 1997, the United States has issued inflation-indexed bonds. By extracting observations about expected inflation and the real interest rate in this market, several studies have found evidence about how real and nominal interest rates react to monetary policy surprises. For example, Gürkaynak, Sack, and Swanson (2003) show that the implied forward 1-year rate at the 9-year horizon responds significantly to a surprise in the federal funds market. They find that the surprise is contained in the expected inflation premium and not in the implied forward real rate.

Kliesen and Schmid find a similar result for the 10-year rate (2004a) and for the real rate (2004b), but in their papers, it is not clear what part of the 10-year term structure is responding to the news. There is one drawback to these measures of the real rate and the expected inflation rate: The maturity of these investments is measured in years, and the analysis does not reveal information about the response of the real interest rate

[1] Clark and Kozicki (2004) survey the literature and show that there is a great deal of uncertainty in real-time estimates of the equilibrium real interest rate.

Michelle T. Armesto is senior research associate and William T. Gavin is a vice president and economist at the Federal Reserve Bank of St. Louis.

or expected inflation in the short end of the term structure.

Cornell and French (1986) provide indirect observations on the short end of the term structure by using a measure of the real interest rate extracted from commodity futures prices. They use this measure to gauge the reaction of real interest rates and expected inflation to surprises in the weekly money supply announcements between October 6, 1977, and March 23, 1984. Their results were somewhat surprising: They found that it was expected inflation in commodity prices and not real returns that went up when there was an unexpected increase in the money supply. These results were obtained using data after October 6, 1979, an era in which the Treasury bill (T-bill) rate responded strongly and positively to surprise increases in the money supply. Before this result, previous authors concluded that these increases in the T-bill rate were due to rising real interest rates—a liquidity effect, perhaps associated with sticky prices (Roley and Walsh, 1985) or with rationing in the market for borrowed reserves (Gavin and Karamouzis, 1985).

We find results reminiscent of Cornell and French (1986). We estimate the market's reaction to surprises in the Fed's interest rate target. The next few sections explain how the market variables and the policy target surprises are constructed. In the results section, we show that expected inflation responds positively and significantly to surprises in the federal funds rate target in the horizon from 0 to 9 months. We also show that there is no significant response of real interest rates out to a year on the term structure. Although the real interest rate and expected inflation rate constructed using averages from commodity futures data are highly volatile, limited information estimates such as the median change can substantially reduce the noise in such measures.

MEASURING THE REAL INTEREST RATE

The real own rate of return for each commodity is implied by the term structure of interest rates in the market for T-bills and the term structure of futures prices in the commodity futures market.[2] Suppose that there is a fixed bundle of consumption goods, Q, that is traded in futures markets in every period. The real interest rate in terms of Q from time t to time $t+k$ is defined as the rate at which the household can contract today to exchange units of Q at time t for units of Q at $t+k$. The cost of bundle Q at time t is S_t defined as

$$(1) \qquad S_t = \sum_{i=1}^{N} q_i S_t^i ,$$

where there are N commodities indexed by i, q_i is the amount of good i in the bundle, and S_t^i is the price of good i at time t.

If there were complete futures markets for all the goods in Q, then the household could contract to buy the bundle at time t for consumption at time $t+k$. The cost of the bundle is given by the futures price, ${}_tF_{t+k}$, which is a sum of individual futures prices

$$(2) \qquad {}_tF_{t+k} = \sum_{i=1}^{N} q_i \, {}_tF_{t+k}^i ,$$

where ${}_tF_{t+k}^i$ is the futures price of good i at time t for delivery k periods ahead. At time t the household can purchase discount bonds that mature at $t+k$ and use the funds to buy the bundle of commodities, Q. The price of the bundle at time t then is ${}_tF_{t+k}\,{}_tB_{t+k}$, where ${}_tB_{t+k}$ is the price of a discount bond that pays one dollar at $t+k$. The gross interest rate from t to $t+k$, $1+{}_tr_{t+k}$, is the ratio of the cost of the bundle today to the cost of the future bundle today:

$$(3) \qquad 1+{}_tr_{t+k} = \frac{S_t}{{}_tF_{t+k}\;{}_tB_{t+k}} .$$

Cornell and French (1986) show that this real rate is an expenditure-weighted average of the commodity own rates.[3]

In the empirical application described here

[2] This section draws heavily from Cornell and French (1986).

[3] They assume that the commodity bundle includes all goods traded in the market, so the commodity own rate is the relevant real interest rate. To the extent that some goods are excluded, this real rate may differ from the real rate that affects consumer and business spending decisions. Gorton and Rouwenhorst (2004) show systematic differences between returns in commodity futures and bonds at lower frequencies.

later, date t is the day of the policy action. We use the maturing contract as the spot price to ensure that the spot and futures prices refer to exactly the same item being traded. However, if the maturing contract does not mature within the next 46 days, we omit this spot price from our data set. We construct a set of futures prices for commodity contracts maturing at k days for $k = \{90, 180, 270, 360\}$. We then calculate the implied forward own rates. When there are no spot prices available, we are able to calculate the implied forward rates because the spot price drops out of the formula. For example, the own rate over the horizon $t + k + 90$ is given by

$$(4) \qquad 1 + {}_t r_{t+k+90} = \frac{S_t}{{}_t F_{t+k+90} \; {}_t B_{t+k+90}},$$

and the implied forward rate from $t + k$ to $t + k + 90$ is given by

$$(5) \qquad 1 + {}_{t+k} r_{t+k+90} = \frac{{}_t F_{t+k} \; {}_t B_{t+k}}{{}_t F_{t+k+90} \; {}_t B_{t+k+90}},$$

where the implied forward rate is the rate at which one can trade the bundle at $t + k$ for the bundle at $t + k + 90$.

MEASURING THE EXPECTED INFLATION RATE

The expected inflation rate in commodities is calculated from the relative bases in commodity markets. The basis is defined as the difference between the spot price and the futures price of a commodity. Over the horizon $t + k + 90$, this expected inflation rate in a given commodity, then, is given as the relative basis, ${}_t b_{t+k+90}$, which is just the basis divided by the spot price:

$$(6) \qquad {}_t b_{t+k+90} = \frac{{}_t F_{t+k+90} - S_t}{S_t},$$

or, in gross terms,

$$(7) \qquad 1 + {}_t b_{t+k+90} = \frac{{}_t F_{t+k+90}}{S_t}.$$

We aggregate relative bases across the commodity bundle to get the expected inflation rate for commodities. We calculate the term structure of implied expected inflation rates in the same manner as we calculated the implied own rates. For example, the implied forward expected inflation rate from $t + k$ to $t + k + 90$ is given by

$$(8) \qquad 1 + {}_{t+k} b_{t+k+90} = \frac{{}_t F_{t+k+90}}{{}_t F_{t+k}},$$

where the implied forward expected inflation rate is the 3-month inflation rate expected for the period $t + k$ through $t + k + 90$. Note that this calculation includes the basis risk—that is, a premium for bearing risk that the actual price in the future will be different from today's futures price. Our web-based data appendix (which appears with this article at http://research.stlouisfed.org/publications/review/) includes summary statistics for the changes in commodity own rates and commodity bases around Fed policy surprises.

THE COMMODITY DATA

The 34 commodities included in our futures market data are listed in Table 1.[4] The commodities were traded on several different North American exchanges: the Coffee Sugar Cocoa Exchange, the Chicago Board of Trade, the Chicago Mercantile Exchange, the New York Mercantile Exchange, the Kansas City Board of Trade, the Minneapolis Grain Exchange, the New York Cotton Exchange, and the Winnipeg Stock Exchange. On average, commodity prices fell 2/3 of a percentage point per year during our sample period. But there was substantial dispersion across commodities. At the high end, palladium was an outlier, rising on average 11.93 percent per year. At the low end, orange juice fell on average 5.79 percent per year. There were contracts expiring for all of the commodities except high-grade copper throughout our full sample period. The first futures contract in high-grade copper expired in 1989.

[4] We chose not to update the data set to include the post-2001 data. Our data, which come directly from the market electronic feeds, were purchased from the Institute for Financial Markets in August 2002. This company, sold to MJK Associates, no longer provides data in the format used in 2002.

Table 1

Commodities Included in the Sample

Commodity	Traded in market	No. of contracts expiring in 1988	No. of contracts expiring in 2001	Average inflation rate
Coffee (KC)	CSCE	5	5	–5.64%
Corn (CN)	CBT	5	7	–1.80%
Feeder cattle (FC)	CME	8	8	0.22%
Gold (GC)	NYMEX	6	12	–3.31%
High-grade copper (HG)	NYMEX	0	12	–5.01%
Live hogs (LH)	CME	7	7	3.53%
Live cattle (LC)	CME	6	6	–0.20%
Oats (OA)	CBT	5	5	–4.71%
Platinum (PL)	NYMEX	4	4	0.26%
Pork bellies (PB)	CME	5	5	5.46%
Silver (SI)	NYMEX	6	12	–2.61%
Soybeans (SY)	CBT	7	7	–4.36%
Soybean meal (SM)	CBT	8	8	–3.64%
Soybean oil (BO)	CBT	8	8	–2.62%
Wheat (WC)	CBT	5	5	–3.49%
Silver 1000 oz (AG)	CBT	6	12	–2.67%
Gold – kilo (KI)	CBT	6	6	–3.64%
Sugar (SB)	CSCE	5	4	–0.20%
Wheat (KW)	KCBT	5	5	–1.56%
Wheat – white (MW)	MGE	5	5	–1.20%
Cotton (CT)	NYCE	5	5	–1.60%
Crude oil – light (CL)	NYMEX	12	12	4.28%
Heating oil (HO)	NYMEX	12	12	2.93%
Liquid propane (PN)	NYMEX	12	12	6.85%
Palladium (PA)	NYMEX	4	4	11.93%
Unleaded gasoline (HU)	NYMEX	12	12	6.45%
Cocoa (CC)	CSCE	5	5	–3.15%
Orange juice (JO)	NYCE	6	6	–5.79%
Rice (NR)	CBT	5	6	–1.44%
Lumber (LB)	CME	6	6	2.85%
Flax seed (WF)	WINN	3	5	–2.76%
Oats (WO)	WINN	5	5	–3.50%
Rapeseed (WP)	WINN	5	7	–1.49%
Wheat (WW)	WINN	5	5	0.04%

NOTE: CSCE, Coffee Sugar Cocoa Exchange; CBT, Chicago Board of Trade; CME, Chicago Mercantile Exchange; NYMEX, New York Mercantile Exchange; KCBT, Kansas City Board of Trade; MGE, Minneapolis Grain Exchange; NYCE, New York Cotton Exchange; WINN, Winnipeg Stock Exchange.

The initial format of the data consisted of hundreds of individual files, each containing detailed information about a particular commodity futures contract: the underlying commodity; the month and year the contract expired; and the open, close, high, and low prices for each day that the contract was traded. From these files we construct time series of prices for all 34 commodities for days surrounding our measure of monetary policy surprises. The difference between the close price on the day before the Fed's target change and the opening price on the day after the Fed's target change is used to gauge how the market responds to incoming information about monetary policy.[5] A term structure of prices for each commodity is constructed using contracts with different expiration dates.

The number and length of contracts varies across commodities. The most frequently traded commodities are crude oil, heating oil, liquid propane, and unleaded gasoline, which have contracts expiring in every month. Others were not as active. Contracts for platinum and palladium expire just four times per year. Additionally, the length of the individual contracts varied. For example, the majority of coffee's commodity contracts (92 percent) were traded for longer than one year, while no contracts were traded for less than one month. Conversely, the majority of flax seed contracts traded within the horizon of 6 to 12 months (86 percent), whereas only two contracts (3 percent) traded for longer than one year. Silver (1000 oz) had the largest quantity of contracts that traded for less than one month (7 percent).

There would be many gaps in the time series of term structures if we insisted on using only those contracts that expired exactly 3, 6, 9, and 12 months from the day of a monetary policy surprise. We use simple decision rules to construct a term structure of prices that approximates prices at spot, 3-month, 6-month, 9-month, and 12-month horizons. First, we looked at all contracts that had been traded in the individual commodity on the day before and day after a monetary policy surprise. We used the maturing contract as a measure of the spot price to ensure that the underlying commodity for the spot price was the same as for the futures contracts. A contract price was considered to be the spot price if the contract had less than 47 days from the day before the policy surprise until expiration. For the 3-month price, a contract was selected if it had between 48 and 137 days until expiration. Similar windows were constructed for the 6-, 9-, and 12-month horizons. For the 6-month futures price, the window was from 138 to 227 days. The 9-month window was from 228 to 317 days, and the 12-month window was from 318 to 417 days until expiration.

When there is no contract expiring in a window, then we have no observation for that commodity. When there was more than one contract within the window, we chose the one closest to our ideal term structure; that is, for the spot price, we choose the contract with the closest date to expiration. For all others, the preference was for the center of the window. For example, for the 3-month futures price for the day after the Fed's policy change, the contract used to represent the 3-month futures price was the one closest to expiring in 90 days. This selection method is similar for the 6-, 9-, and 12-month futures price. These prices were then used to compute our own rates and bases.

WHAT IS A SURPRISE IN THE FEDERAL FUNDS RATE TARGET?

We use measures of the monetary policy surprise as constructed by Poole, Rasche, and Thornton (2002). Data from the federal funds futures market is used to measure the expected change in monetary policy. The federal funds futures market is a bet on the average effective federal funds rate for the month in which the contract matures; as a result, it is an estimate of market expectations about the average level of the federal funds rate for that month.

The Chicago Board of Trade began trading federal funds futures contracts in October 1988. When this trading began, the Federal Reserve was using a target or "expected trading level" for the

[5] Note that, since 1994, the announcements of policy changes are scheduled for release at a preannounced time, so that today one could use higher-frequency data and a much smaller window in which to measure market reaction. But in the period before 1994, we do not have good information about when, during the day, the market learned about the policy change.

federal funds rate as a guide for daily open market operations, but it did not announce its short-run targets until 1994. Poole, Rasche, and Thornton describe the history of Fed policy changes as embodied in a target or expected trading level for the federal funds rate for the pre-1994 period; they also show that the market often could see when the Fed's target was changed. Fed policy has been transparent since 1994, but what the market expects as it compares with what actually occurs is still a relevant issue.

The unexpected component of the Fed's actions is implied by the change in the futures market price from the day before to the day after the policy change. Suppose fff_t^h denotes the rate on the h-month federal funds futures contract on day t. This rate is equivalent to the sum of the expectation on day t of the federal funds rate on each day of the month, averaged across the length of the month. Hence,

$$(9) \qquad fff_t^h = 1/m \sum_{i=1}^{m} E_t(ff_i^h),$$

where ff_i^h denotes the federal funds rate on day i of the hth month, E_t denotes the expectation on day t, and m denotes the number of days in the month.

Next, consider if the Fed successfully targets the federal funds rate, so that the actual federal funds rate is equal to the target plus an i.i.d. mean zero error term:

$$(10) \qquad ff_t = ff_i^* + \eta_t .$$

The expectation of the future federal funds rate depends on expectations about the policy target. The surprise in the federal funds market is calculated as the change in the federal funds futures price following a monetary policy action. Substituting the federal funds target into the formula for the federal funds futures rate and taking the difference yields

$$(11) \qquad \Delta fff_t^h = 1/m \sum_{i=1}^{m} \left[E_t(ff_i^{*h}) - E_{t-1}(ff_i^{*h}) \right].$$

If the policy change was perfectly anticipated, then there will be no change in the futures price.

Before 1994, whether or not the market was aware of the Fed's actions is an issue. Poole, Rasche, and Thornton (2002) use reports in the financial press to distinguish between days when the market was aware and days when they were unaware that policy had changed. We examine only days in which they find that the market was aware that policy had changed.

MEASURING THE MARKET RESPONSE

The market response following a policy action is measured as the change in the own rate or expected inflation calculated using the closing price from the day before the policy change to the opening price on the day after. The expiration date is fixed, so the actual horizon gets 2 days shorter during the interval of the policy change. The change in the 3-month-ahead own rate for a given commodity is $\Delta_0 r_{90,t}^i = {}_0r_{90,t+1}^i - {}_0r_{90,t-1}^i$. The change in the 9-month-ahead implied forward 3-month rate is $\Delta_{270} r_{360,t}^i = {}_{270}r_{360,t+1}^i - {}_{270}r_{360,t-1}^i$, and so on for the other own rates and expected inflation rates. We calculate the average change in the implied forward commodity own rate for the bundle as $\Delta_k r_{k+90,t}$ and the average change in expected inflation as $\Delta_k b_{k+90,t+1}$.[6] We also calculate the aggregate commodity own rates and expected inflation rates as the median commodity change. The median changes are designated with a *med* superscript. For example, the change in the 9-month-ahead 3-month expected inflation rate from the day before the monetary policy surprise to the day after is given as $\Delta_{270} b_{360,t}^{med}$.

The volatility of the implied forward T-bill rates, commodity own rates, and commodity expected inflation rates following a monetary policy surprise are shown in Table 2. In the first row, we report summary statistics about the change in the implied forward 3-month T-bill rates following surprises in the federal funds rate. At all four horizons, the standard deviation of the change is 10 or 11 basis points. In the second and third rows we report the volatility of changes in the commodity own rates. As Mishkin (1990) notes, this series is quite volatile relative to inter-

6 Cornell and French (1986) weighted the individual commodities by the inverse of measures of volatility in each market. Mishkin (1990) used an unweighted average.

Table 2

Volatility of Changes in Rates Following Federal Funds Rate Surprises

	0 to 3 months	3 to 6 months	6 to 9 months	9 to 12 months
Implied forward T-bill rates	10	11	11	11
Commodity own rates				
Mean commodity	94	58	43	58
Median commodity	46	24	19	23
Commodity inflation rates				
Mean commodity	94	57	42	55
Median commodity	46	24	18	19

NOTE: Standard deviations in basis points at annual rates.

est rates: The standard deviation of changes following surprises in the federal funds rate ranges from a high of 94 basis points in the near term to 43 basis points for the implied own rate in the 6- to 9-month horizon. In every case, the standard deviations of the median changes are less than half the standard deviations of the mean changes. Here we find that the standard deviations of changes in commodity own rates ranged from 46 basis points in the near horizon to 19 points for the implied own rates in the 6- to 9-month horizon.

The fourth and fifth rows report the standard deviations of the commodity expected inflation rates. Here the volatility pattern is very similar to the pattern for the own rates. The standard deviation of changes following surprises in the federal funds rate ranges from a high of 94 basis points in the near term to 42 basis points for the implied own rate in the 6- to 9-month horizon. Also, the standard deviations of the median changes are less than half the standard deviations of the mean changes. Here we find that the standard deviations of changes in commodity expected inflation rates ranged from 46 basis points in the near horizon to 18 points for the implied own rate in the 6- to 9-month horizon.

RESULTS

We run simple regressions relating changes in the mean and median commodity own rates to surprises in the policy action. The rationale for this model is simply that at the moment before the policy change, market prices reflect all the information that is relevant, including expectations about policy. The relationship between changes in the mean and median own rates and the policy surprise is measured using the following regressions:

$$\text{(12)} \qquad \Delta_k r_{k+90,t} = \alpha_1^r + \beta_1^r \Delta fff_t^{\,h} + \varepsilon_{1,t}^r \text{ and}$$

$$\text{(13)} \qquad \Delta_k r_{k+90,t}^{med} = \alpha_2^r + \beta_2^r \Delta fff_t^{\,h} + \varepsilon_{2,t}^r ,$$

for $k = 0, 90, 180$, and 270. Similar regressions are used to measure the response of the mean and median commodity expected inflation rates to a monetary policy surprise:

$$\text{(14)} \qquad \Delta_k b_{k+90,t} = \alpha_1^b + \beta_1^b \Delta fff_t^{\,b} + \varepsilon_{1,t}^b \text{ and}$$

$$\text{(15)} \qquad \Delta_k b_{k+90,t}^{med} = \alpha_2^b + \beta_2^b \Delta fff_t^{\,h} + \varepsilon_{2,t}^b .$$

The estimation results for these equations are reported in Table 3. The important effect of idiosyncratic shocks results in very low R^2s for the regression of the commodity own rates and expected inflation rates on the federal funds surprise.[7] In the top half we see that in no case is the

[7] Poole, Rasche, and Thornton (2002) note that the surprise in the federal funds target is measured with error, leading to biased estimates of the coefficient on the federal funds surprise. The measurement error biases the coefficient toward zero. They show that the variance of the measurement error is small and is unlikely to affect the qualitative nature of our results.

Table 3

Regressions of Commodity Own Rates and Basis Changes on the Unexpected Component of Federal Funds Announcements

	β	*t*-Statistic	SEE	R^2
Commodity own rates				
$\Delta_0 r_{90,t+1}$	–0.65	–0.62	0.96	0.00
$\Delta_0 r^{med}_{90,t+1}$	–0.40	–0.78	0.46	0.01
$\Delta_{90} r_{180,t+1}$	–0.03	–0.04	0.54	0.00
$\Delta_{90} r^{med}_{180,t+1}$	0.18	0.68	0.24	0.01
$\Delta_{180} r_{270,t+1}$	–0.41	–0.96	0.40	0.01
$\Delta_{180} r^{med}_{270,t+1}$	0.02	0.08	0.18	0.00
$\Delta_{270} r_{360,t+1}$	0.73	1.25	0.54	0.02
$\Delta_{270} r^{med}_{360,t+1}$	0.29	1.23	0.21	0.02
Commodity inflation rates				
$\Delta_0 b_{90,t+1}$	1.36	1.31	0.95	0.02
$\Delta_0 b^{med}_{90,t+1}$	1.10	**2.23**	0.45	0.06
$\Delta_{90} b_{180,t+1}$	0.79	1.38	0.52	0.02
$\Delta_{90} b^{med}_{180,t+1}$	0.58	**2.28**	0.23	0.06
$\Delta_{180} b_{270,t+1}$	1.11	**2.61**	0.39	0.07
$\Delta_{180} b^{med}_{270,t+1}$	0.68	**3.84**	0.16	0.15
$\Delta_{270} b_{360,t+1}$	–0.12	–0.22	0.51	0.00
$\Delta_{270} b^{med}_{360,t+1}$	0.32	1.63	0.18	0.03

NOTE: Bold indicates that the *t*-statistic is significant at the 5 percent critical level. SEE is the standard error of the equation.

response of the own rate significantly different from zero. At the 9- to 12-month horizon, the response of the real rate is positive and relatively large, but not statistically significant.

The bottom half reports the results for the commodity price expected inflation. The results for the mean response are not significant except in the case of the implied 3-month expected inflation rate from 6 to 9 months. We get significant results more often when we use the median response. Mishkin (1990) argued that the Cornell and French results are unreliable indicators of real interest rate behavior because commodity market returns are so highly variable relative to expected inflation and nominal interest rates that there is little signal about real interest rates in the data. We reduce volatility by using the median response measure, which increases reliability in the significance of our results.

Using the median measure of commodity price expected inflation, we find that expected inflation responded positively and significantly to surprises in the federal funds target for the first three horizons. Only in the case of the 9- to 12-month rate is the response not statistically significant at the 5 percent level. Relative to the mean price or own rate, the median appears to effectively filter a substantial amount of idiosyncratic noise from the commodity futures data.

DISCUSSION

The results in Table 2 appear to be at odds with Gürkaynak, Sack, and Swanson (2003) and Kliesen and Schmid (2004a), who report that inflation expectations decline when there is a surprise increase in the federal funds target. The different results, however, refer to different points on the term structure. We find that the positive impact is largest at the 3-month horizon and is not statistically significant at the 1-year horizon.

It is also quite possible that the high-frequency response of commodity prices is not the same as the response of prices across the broad spectrum of goods in the economy. However, Gorton and Rouwenhorst (2004) use monthly data to show that nominal commodity returns are highly and positively correlated with CPI inflation, both its expected and unexpected components.

Our results appear to be at odds with conventional wisdom as well, which suggests that inflation responds to policy actions only after a long lag. The conventional view implies that the immediate response of interest rates to a monetary policy shock is by the real component. Recent developments in macroeconomic theory focus on policy in general equilibrium models. To capture this conventional wisdom, economists have used New Keynesian models that incorporate some form of price stickiness.[8] Gavin, Keen, and Pakko (2004) analyze the effects of monetary policy shocks in such a model where Calvo-style pricing means that prices change on average once per year and the central bank uses an interest rate rule to implement policy. Within this framework, there is an important difference in the effect of a shock to the interest rate depending on whether the shock is perceived to be temporary or persistent.[9] Transitory policy shocks affect both expected inflation and the real interest rate. With a transitory shock, the federal funds rate returns to its original level within a few quarters and the implied future short-term rate is essentially unaffected by the shock at the 4-quarter horizon.

However, if the shock is persistent—that is, the market expects that a change in the federal funds target is likely to be fairly permanent—then the effect of a positive shock on the implied future short-term rate may be positive for many quarters. But almost all the effect is due to higher expected inflation. The predicted real interest rate effects of a persistent shock to the federal funds rate target are an order of magnitude smaller than are the effects following a transitory shock. That is, state-of-the-art macroeconomic theory predicts that the real interest rate will not respond to federal funds rate target shocks that are highly persistent.

It is important to understand, then, whether monetary policy shocks are perceived as relatively transitory or relatively permanent. Many empirical studies find that the level of the federal funds rate behaves as if it has a random walk component. Using the futures market data to derive the shocks as we do, Faust, Swanson, and Wright (2004) report that shocks to the change in the federal funds rate are highly persistent.

In the top panel of Table 4, we report the response of the term structure of implied future 3-month T-bill rates to surprises in monetary policy. These coefficients were calculated by regressing the change in the implied forward T-bill rates on the monetary policy surprises in the following regression:

$$\Delta_k Tbill_{k+90,t} = \alpha_1 + \beta_1 \Delta fff_t^h + \varepsilon_{1,t}, \tag{16}$$

where $\Delta_k Tbill_{k+90,t}$ is the change in the implied forward T-bill rate for values of k equal to 0, 90, 180, and 270. We find that for the first four quarters, the response in the implied forward T-bill rate is fairly equal across the term structure from 3 months to 12 months. A 10-basis-point surprise in the federal funds rate leads to a 7-basis-point rise in the 3-month rate and to a pattern of 8-, 7-, and 6-basis-point increases in the implied future 3-month rates at horizons ending in 6, 9, and 12 months, respectively. In the bottom panel, we report similar regressions in which the dependent variables are changes in the Treasury bill rates, $\Delta_0 Tbill_{k+90,t}$, maturing in $k + 90$ days. The regression is given as

$$\Delta_0 Tbill_{k+90,t} = \alpha_2 + \beta_2 \Delta fff_t^h + \varepsilon_{2,t}. \tag{17}$$

The results show that the 12-month rate changes by only very slightly less than the 3-month rate.

Markets appear to expect that a shock to the federal funds target is relatively permanent—at least that it persists intact for the first year. Accord-

[8] See Woodford (2003, Chap. 3) for a comprehensive analysis of the New Keynesian model.

[9] In that model the shocks follow an AR(1) process. The transitory shock has an AR(1) parameter equal to 0.3 and the persistent shock has a parameter value of 0.95.

Table 4

Response of T-Bill Rates to Federal Funds Target Surprises

Maturity	β	*t*-Statistic	SEE	R^2
Implied forward rates				
$\Delta_0 Tbill_{90,t+1}$	0.71	**9.89**	0.07	0.54
$\Delta_{90} Tbill_{180,t+1}$	0.77	**9.75**	0.07	0.53
$\Delta_{180} Tbill_{270,t+1}$	0.69	**7.61**	0.08	0.41
$\Delta_{270} Tbill_{360,t+1}$	0.61	**5.89**	0.09	0.29
T-bill rates				
$\Delta_0 Tbill_{90,t+1}$	0.71	**9.89**	0.07	0.53
$\Delta_{90} Tbill_{180,t+1}$	0.73	**11.04**	0.06	0.58
$\Delta_{180} Tbill_{270,t+1}$	0.73	**10.51**	0.06	0.56
$\Delta_{270} Tbill_{360,t+1}$	0.69	**9.12**	0.07	0.49

NOTE: Bold indicates that the *t*-statistic is significant at the 5 percent critical level. SEE is the standard error of the equation.

ing to general equilibrium macro theory, the shock to the federal funds rate should have significant effects on expected inflation, but not on real interest rates, which is what we see in the commodity futures market.

CONCLUSION

Although the commodity futures data contain a substantial amount of idiosyncratic noise, they remain an important source of information about how markets respond to the implementation of monetary policy. Evidence presented in this paper shows that real rates of return in commodity markets do not appear to react to surprises in the federal funds target. This result complements research showing that real rates in the long-term market for inflation-protected Treasury securities also do not respond to these surprises. This is also the result predicted by New Keynesian macroeconomic models if the shocks to the federal funds target appear to be persistent. Persistent shocks lead to a significant (almost one-for-one) response by expected inflation but no measurable response by the real interest rate.

Our results show that despite the relative noise in commodity futures markets, the commodity expected inflation rate does respond significantly to surprises in the federal funds rate. The result is consistent with modern theory, but not with conventional wisdom. Conventional wisdom suggests that expected inflation does not respond immediately, but only with a long lag—and the response should be negative, as was found in the long-term indexed bond market. Gürkaynak, Sack, and Swanson (2003) and Kliesen and Schmid (2004a) find that the long-term expected inflation rate falls when there is a surprise increase in the federal funds rate. Our results suggest that the short-term response is different. That is, expected inflation, at least as observed in commodity markets, over the next 3 to 9 months moves in the same direction as a surprise in the federal funds rate target.

REFERENCES

Clark, Todd E. and Kozicki, Sharon. "Estimating Equilibrium Real Interest Rates in Real Time." Research Working Paper P04-08, Federal Reserve Bank of Kansas City, September 2004.

Cornell, Bradford and French, Kenneth R. "Commodity Own Rates, Real Interest Rates, and Money Supply Announcements." *Journal of Monetary Economics*, 1986, *18*, pp. 3-20.

Faust, Jon; Swanson, Eric T. and Wright, Jonathan H. "Identifying VARS Based on High Frequency Futures Data." *Journal of Monetary Economics*, 2004, *51*, pp. 1107-31.

Gavin, William T. and Karamouzis, Nicholas V. "The Reserve Market and the Information Content of M1 Announcements." Federal Reserve Bank of Cleveland *Review*, 1985, Quarter I, pp. 11-28.

Gavin, William T.; Keen, Benjamin D. and Pakko, Michael R. "The Monetary Instrument Matters." Working Paper 2004-026A, Federal Reserve Bank of St. Louis, November 2004.

Gorton, Gary and Rouwenhorst, K. Geert. "Facts and Fantasies About Commodity Futures." NBER Working Paper 10595, National Bureau of Economic Research, June 2004.

Gürkaynak, Refet S.; Sack, Brian and Swanson, Eric. "The Excess Sensitivity of Long-Term Interest Rates: Evidence and Implications for Macroeconomic Models." Board of Governors of the Federal Reserve System *Finance and Economics Discussion Series* 2003-50, August 13, 2003.

Kliesen, Kevin L. and Schmid, Frank A. "Monetary Policy Actions, Macroeconomic Data Releases, and Inflation Expectation." Federal Reserve Bank of St. Louis *Review*, May/June 2004a, *86*(3), pp. 9-21.

Kliesen, Kevin L. and Schmid, Frank A. "Do Productivity Growth, Budget Deficits, and Monetary Policy Actions Affect Real Interest Rates? Evidence From Macroeconomic Announcement Data." Working Paper 2004-019A, Federal Reserve Bank of St. Louis, September 2004b.

Kuttner, Kenneth N. "Monetary Policy Surprises and Interest Rates: Evidence from the Federal Funds Futures Market." *Journal of Monetary Economics*, June 2001, *47*(3), pp. 523-44.

Mishkin, Frederic S. "Can Futures Market Data Be Used to Understand the Behavior of Real Interest Rates?" *Journal of Finance*, March 1990, *45*, pp. 245-57.

Poole, William; Rasche, Robert H. and Thornton, Daniel L. "Market Anticipations of Monetary Policy Actions." Federal Reserve Bank of St. Louis *Review*, July/August 2002, *84*(4), pp. 65-93.

Roley, V. Vance and Walsh, Carl E. "Monetary Policy Regimes, Expected Inflation, and the Response of Interest Rates to Money Announcements." *Quarterly Journal of Economics*, 1985, *100*(5, Supplement), pp. 1011-37.

Woodford, Michael. *Interest and Prices: Foundations of a Theory of Monetary Policy*. Princeton: Princeton University Press, 2003.

Using Implied Volatility to Measure Uncertainty About Interest Rates

Christopher J. Neely

Option prices can be used to infer the level of uncertainty about future asset prices. The first two parts of this article explain such measures (implied volatility) and how they can differ from the market's true expectation of uncertainty. The third then estimates the implied volatility of three-month eurodollar interest rates from 1985 to 2001 and evaluates its ability to predict realized volatility. Implied volatility shows that uncertainty about short-term interest rates has been falling for almost 20 years, as the levels of interest rates and inflation have fallen. And changes in implied volatility are usually coincident with major news about the stock market, the real economy, and monetary policy.

Federal Reserve Bank of St. Louis *Review*, May/June 2005, *87*(3), pp. 407-25.

Economists often use asset prices along with models of their determination to derive financial markets' expectations of events. For example, monetary economists use federal funds futures prices to measure expectations of interest rates (Krueger and Kuttner, 1995; Pakko and Wheelock, 1996). Similarly, a large literature on fixed and target zone exchange rates has used forward exchange rates to measure the credibility of exchange rate regimes or to predict their collapse (Svensson, 1991; Rose and Svensson, 1991, 1993; Neely, 1994).

But it is often helpful to gauge the *uncertainty* associated with future asset prices as well as their expectation. Because option prices depend on the perceived volatility of the underlying asset, they can be used to quantify the expected volatility of an asset price (Latane and Rendleman, 1976). Such estimates of volatility, called implied volatility (IV), require some heroic assumptions about the stochastic (random) process governing the underlying asset price. But the usual assumptions seem to provide very reasonable forecasts of volatility. That is, IV is a highly significant but biased predictor of volatility, which often encompasses other forecasts.

Readers who are already familiar with the basics of options might wish to skip the first section of this article; it explains how option prices are determined by the cost of a portfolio of assets that can be dynamically traded to provide the option payoff. Readers who are unfamiliar with options might wish to start with the glossary of option terms at the end of this article and the insert on the basics of options (boxed insert 1). The second section reviews the relation between IV and future volatility, showing how option pricing formulas can be "inverted" to estimate volatility. The third section measures the IV of short-term interest rates over time and discusses how such measures can aid in interpreting economic events.

HOW DOES ONE PRICE OPTIONS?

Options are a *derivative* asset. That is, option payoffs depend on the price of the underlying asset. Because of this, one can often exactly replicate the payoff to an option with a suitably

Christopher J. Neely is a research officer at the Federal Reserve Bank of St. Louis. Joshua Ulrich provided research assistance.

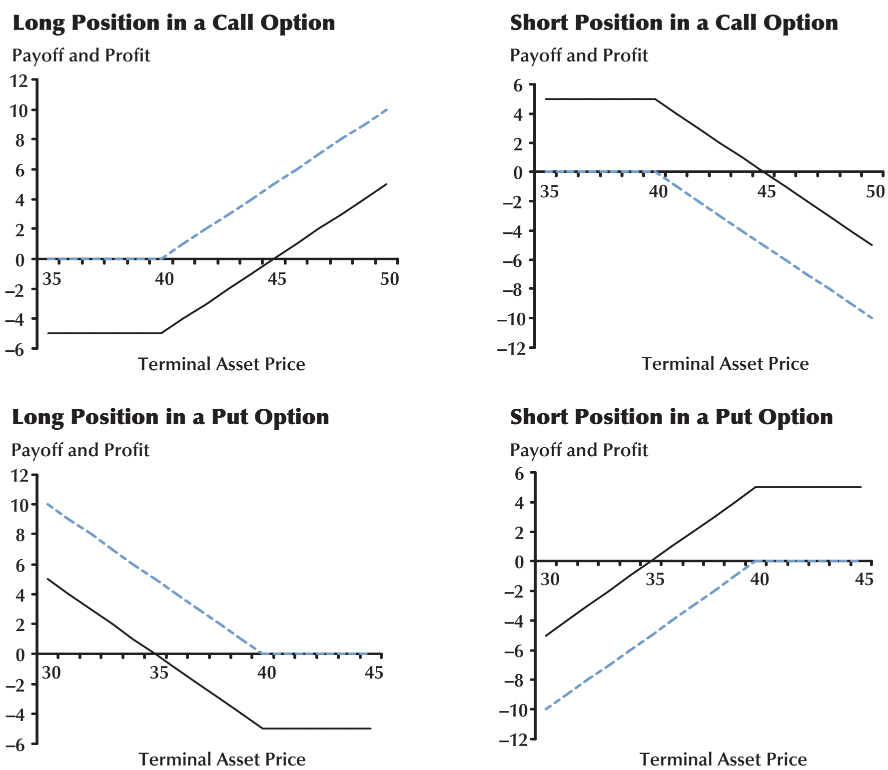
BOXED INSERT 1: OPTION BASICS

A call is an option to buy an underlying asset; a put is an option to sell the underlying asset. A European option can be exercised only at the end of its life; an American option can be exercised at any time prior to expiry.

One can either buy or sell options. In other words, one can be long or short in call options or long or short in put options. The payoff to a long position in a European call option with a strike price of X is $max(S_T - X, 0)$. The payoff to a long position in a European put option with a strike price of X is $max(X - S_T, 0)$. The payoffs to short positions are the negatives of these. The figure below shows the payoffs to the four option positions as a function of the terminal asset price for strike prices of \$40.

The relation of the current price of the underlying asset to the strike price of an option defines the option's "moneyness." Options that would net a profit if they could be exercised immediately are said to be "in the money." Options that would lose money if they were exercised immediately are "out of the money," and those that would just break even are "at the money." For example, if the underlying asset price is \$50, then a call option with a strike price of \$40 is in the money, while a put option with the same strike would be out of the money.

Because the holder of an option has limited risk from adverse price movements, greater asset price volatility tends to raise the price of an option. Because the uncertainty about the future asset price generally increases with time to expiry, options generally have "time value," meaning that—all else equal—American options with greater time to expiry will be worth more.[1]



NOTE: The four figures display the payoffs (blue dashed line) and the profits (black line) for the four option positions as a function of the terminal asset price.

[1] European options on equities can have negative time value in the presence of dividends.

managed portfolio of the underlying asset and a riskless asset. The set of assets that replicates the option payoff is called the *replicating portfolio*. This section explains how arbitrage equalizes the price of the option and the price of the replicating portfolio.

Pricing an Option with a Binomial Tree

A simple numerical example will help explain how the price of an option is equal to the price of a portfolio of assets that can replicate the option payoff. Suppose that a stock price is currently $10 and that it will either be $12 or $8 in one year.[1] Suppose further that interest rates are currently 5 percent. A one-year *European call* option with a strike price of $10 gives the buyer the right, but not the obligation, to purchase the stock for $10 at the end of one year.[2] If the stock price goes up to $12, the option will be worth $2 because it confers the right to pay 10 for an asset with a $12 market price. But if the stock price falls to $8, the option will be worthless because no one would want to buy a stock at the strike price when the market price is lower.

Suppose that the First Bank of Des Peres (FBDP) sells one call option on one share of a non-dividend-paying stock and simultaneously buys some amount, call it Δ, shares of the stock. If the stock price goes up to $12, the FBDP's portfolio will be worth the value of its stock, less the value of the option: $\$12\Delta - \$2$. If the stock price falls to $8, the option will be worthless and the FBDP's portfolio will only be worth $\$8\Delta$. The key to option pricing is that the FBDP can choose Δ to make the value of its portfolio the same in either state of the world: It chooses $\Delta = 1/2$, to make $\$12\Delta - \$2 = \$8\Delta - \$1$. That is, if the FBDP buys $\Delta = 1/2$ units of the stock after selling the call option, it will have a riskless payoff to its portfolio of $4.

Because this payoff is riskless, the portfolio of a short call option and 1/2 share of the stock must earn the riskless return. If it did not, there would be an arbitrage opportunity. The initial cost of the portfolio is the cost of the Δ shares of stock ($\$10\Delta$) less the price of the call option ($C). The initial cost of the portfolio must equal its discounted riskless payoff ($\$4e^{-0.05}$):

$$\$10\Delta - C = \$4e^{-0.05}.^{3} \tag{1}$$

Using the fact that $\Delta = 1/2$, the price of the call option must be

$$C = \$10\frac{1}{2} - \$4e^{-0.05} = \$1.1951. \tag{2}$$

If the price of the call option were more than $1.1951, one could make a riskless profit by selling the option and holding 1/2 shares of the stock.[4] If the call option price were less than $1.1951, one could make an arbitrage profit by buying the call and shorting 1/2 shares of the stock.

An equivalent way to look at the problem is to create the portfolio that replicates the initial investment/payoff of the call option. That is, the FBDP could borrow $5 and buy 1/2 of a share of the stock. At the end of the year, the 1/2 share of stock would be worth either $6 or $4 and the FBDP would owe ($\$5e^{0.05}$=) $5.2564 on the money it borrowed. The initial investment would be zero and the payoff would be $0.7436 in the first state and –$1.2564 in the second state. This is the same initial investment/payoff structure as borrowing $1.1951 and buying the call option with a strike price of $10. In other words, the portfolio that replicates the call option in this example is a 1/2 share of the stock and an equal short position in a riskless bond.

Introductory textbooks on derivatives, like Hull (2002), Jarrow and Turnbull (2000), or Dubofsky and Miller (2003), provide a much more

[1] This example assumes that the stock pays no dividends. If it did pay known dividends, it could be priced in a similar way.

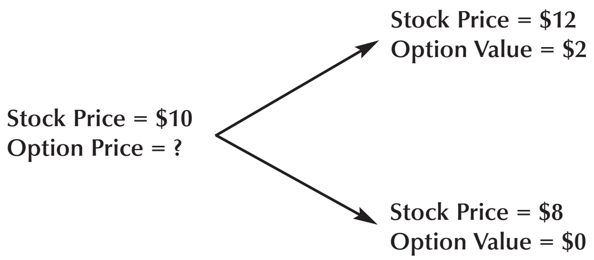
[2] A *European option* confers the right to buy or sell the underlying asset for a given price at the *expiry* of the option. An *American option* can be exercised on or before the expiry date. A *call (put) option* confers the right, but not the obligation, to buy (sell) a particular asset at a given price, called the *strike price*.

[3] If the continuously compounded interest rate is 5 percent, the price of a riskless bond with a one-year payoff of $4 would have a price of $\$4e^{-0.05}$.

[4] Suppose that the call option cost $1.30. One would sell the call option, borrow $3.70, and use the proceeds of the option sale and the borrowed funds to buy 1/2 share of stock. If the first state of the world occurs, the writer of the option will have $6 in stock but will pay $2 to the option buyer and ($3.70e^{0.05}$ =) $3.89 to the bank that loaned him the funds originally. He will make a riskless profit of $0.11. Similarly, in the second state of the world, the option expires worthless and the option writer sells the 1/2 share of stock for $4, pays the loan off with $3.89 and again makes $0.11 riskless profit.

Figure 1

Pricing a Call Option with a Binomial Tree



NOTE: The figure illustrates values that a hypothetical stock could take, along with the value of a call option on that stock with a strike price of $10.

extensive treatment of binomial trees as well as information about how options pricing formulas change for different types of assets.

Black-Scholes Valuation

The preceding example, illustrated in Figure 1, was a one-step binomial tree. The option price was calculated under the assumption that the stock could take one of two known values at expiry. Suppose instead that the stock could move up or down several times before expiration. In this case, one can calculate an option price by computing each possible value of the option at expiry and working backward to get the price at the beginning of the tree. As the asset prices rise and the call option goes "into the money," the replicating portfolio holds more of the underlying asset and less of the riskless bond.[5] At each point in time, the option writer chooses the position in the underlying asset to maintain a riskless payoff to the hedged portfolio—the combination of the positions in the option, the underlying asset, and the riskless bond. The position in the underlying asset is equal to the rate of change in the option value with respect to the underlying asset price. This rate of change is known as the option's "delta" and the continuous process of adjustment of the underlying asset position is known as "delta hedging." The limit of the formula for an option price from an n-step binomial tree, as n goes to infinity, is the Black-Scholes (BS) formula (Black and Scholes, 1972).[6]

The BS formula expresses the value of a European call or put option as a function of the underlying asset price (S), the strike price (X), the interest rate (r), time to expiry (T), and the variance of the underlying asset return (σ^2). Higher asset price volatility means higher option prices because the downside risk is always limited, whereas the upside potential is not. Therefore, option prices increase with expected volatility. The formula for the price of a European call option on a spot asset that pays no dividends or interest is the following:

(3) $$C = S_0 N(d_1) - Xe^{-rT} N(d_2),$$

where $d_1 = \dfrac{\ln(S_0/X) + (r + \sigma^2/2)T}{\sigma\sqrt{T}}$ and

$$d_2 = \frac{\ln(S_0/X) + (r - \sigma^2/2)T}{\sigma\sqrt{T}} = d_1 \sigma\sqrt{T}$$

and $N(*)$ is the cumulative normal density function. Hull (2002), Jarrow and Turnbull (2000), and Dubofsky and Miller (2003) provide formulas for put options and options on other types of assets.

The BS formula strictly applies to European options only—not to American options, which can be exercised any time prior to expiry—and it requires modifications for assets that pay dividends, such as stocks, or that don't require an initial outlay, such as futures.[7] Further, the BS model makes some strong assumptions: that the underlying asset price follows a lognormal random walk, that the riskless rate is a known function of time, that one can continuously adjust one's

[5] A call (put) option is said to be "in the money" if the underlying asset price is greater (less) than the strike price. If the underlying asset price is less (greater) than the strike price, the call (put) option is "out of the money." When the underlying asset price is near (at) the strike price, the option is "near (at) the money."

[6] There are several ways to derive the BS formula that differ in their required assumptions (Merton, 1973b). Wilmott, Howison, and Dewynne (1995) provide a nice introduction to the mathematics of the BS formula and Wilmott (2000) extends that treatment to cover the price of volatility risk. Boyle and Boyle (2001) discuss the history of option pricing formulas.

[7] Black (1976) provides the formula for options on futures, rather than spot assets. Barone-Adesi and Whaley (1987) provide an approximation to the BS formula that accounts for early exercise.

position in the underlying asset (delta hedging), and that there are no transactions costs on the underlying asset and no arbitrage opportunities. Despite these strong assumptions, the BS model is very widely used by practitioners and academics, often fitting the data reasonably well even when its assumptions are clearly violated.

Does IV Predict Realized Volatility?

The BS model expresses the price of a European call or put option (C or P) as a function of five arguments $\{S, X, r, T, \text{and } \sigma^2\}$. Of those six quantities, five are observable as market prices or features of the option contract $\{C, S, X, r, T\}$. The BS formula is frequently inverted to solve for the sixth quantity, the IV $\{\sigma\}$ of log asset returns in terms of the observed quantities. This IV is used to predict the volatility of the asset return to expiry.

Ironically, the BS formula usually used to derive IV assumes that volatility is constant. Hull and White (1987) provide the foundation for the practice of using a constant-volatility model to predict stochastic volatility (SV): If volatility evolves independently of the underlying asset price and no priced risk is associated with the option, the correct price of a European option equals the expectation of the BS formula, evaluating the variance argument at average variance until expiry:

$$\begin{aligned} C(S_t, V_t, t) &= \int_t^T C^{BS}(\bar{V}) h(\bar{V} | \sigma_t^2) d\bar{V} \\ &= E[C^{BS}(\bar{V}_{t,T}) | V_t], \end{aligned} \tag{4}$$

where the average variance until expiry is denoted as

$$\bar{V}_{t,T} = \frac{1}{T-t} \int_\tau^T V_\tau d\tau$$

and its square root is usually referred to as realized volatility (RV).[8]

Bates (1996) points out that the expectation in (4) is taken with respect to variance until expiry, not standard deviation until expiry. Therefore, one cannot use the linearity of the BS formula with respect to standard deviation to justify passing the expectation through the BS formula. That is, one cannot claim that the correct price of a call option under stochastic volatility is the BS price evaluated at the expected value of the standard deviation until expiry. That is, it is *not* true that

$$C\left(S_t, \sqrt{V_t}, t\right) = C^{BS}\left(E\sqrt{\bar{V}}_{t,T} \middle| V_t\right). \tag{5}$$

Instead, Bates (1996) approximates the relation between the BS IV and expected variance until expiry with a Taylor series expansion of the BS price for an at-the-money option. That is, for at-the-money options, the BS formula for futures reduces to

$$C^{BS} = e^{-rT} F\left[2N\left(\frac{1}{2}\sigma\sqrt{T}\right) - 1\right].$$

This can be approximated with a second-order Taylor expansion of $N(*)$ around zero, which yields

$$C^{BS} \approx e^{-rT} F\sigma\sqrt{T/(2\pi)}\ .$$

Another second-order Taylor expansion of that approximation around the expected value of variance until expiry shows that the BS IV is approximately the expected variance until expiry:

$$\hat{\sigma}_{BS}^2 \approx \left(1 - \frac{1}{8}\frac{Var(\bar{V}_{t,T})}{(E_t\bar{V}_{t,T})^2}\right)^2 E_t\bar{V}_{t,T}. \tag{6}$$

That is, the BS-implied variance (σ_{BS}^2) understates the expected variance of the asset until expiry ($E_t\bar{V}_{t,T}$). Similarly, BS-implied standard deviation (σ_{BS}) slightly understates the expected standard deviation of asset returns.[9]

The Volatility Smile

Volatility is constant in the BS model; IV does not vary with the "moneyness" of the option. That is, if the BS model assumptions were literally true, the IV from a deep-in-the-money call should be the same as that from an at-the-money call or an in-the-money put. In reality, for most assets, IV does vary with moneyness. A graph of IV versus moneyness is often referred to as the "volatility

[8] Romano and Touzi (1997) extend the Hull and White (1987) result to include models that permit arbitrary correlation between returns and volatility, like the Heston (1993) model.

[9] Note that (6) depends on (4), which assumes that there is no priced risk associated with holding the option. That is, (6) requires that changes in volatility do not create priced risk for an option writer.

smile" or "volatility smirk," depending on the shape of the relation. Research attributes the volatility smile to deviations from the BS assumptions about the evolution of the underlying asset prices, such as the presence of stochastic volatility, jumps in the price of the underlying asset, and jumps in volatility (Bates, 1996, 2003).

The existence of the volatility smile brings up the question of which strike prices—or combinations of strike prices—to use to compute IV. In practice, IV is usually computed from a few near-the-money options for three reasons (Bates, 1996): (i) The BS formula is most sensitive to IV for at-the-money options. (ii) Near-the-money options are usually the most heavily traded, resulting in smaller pricing errors. (iii) Beckers (1981) showed that IV from at-the-money options provides the best estimates of future realized volatility. While researchers have varied the number and types of options as well as the weighting procedure, it has been common to rely heavily on a few at-the-money options.

Constructing IV from Options Data

At each date, IV is chosen to minimize the unweighted sum of squared deviations of Barone-Adesi and Whaley's (1987) formula for pricing American options on futures with the actual settlement prices for the two nearest-to-the-money call options and two nearest-to-the-money put options for the appropriate futures contract.[10] That is, IV is computed as follows:

(7) $$\sigma_{IV,t,T} = \arg\min_{\sigma_{t,T}} \sum_{i=1}^{4} \left(BAW_i(\sigma_{t,T}) - Pr_{i,t}\right)^2,$$

where $Pr_{i,t}$ is the observed settlement premium (price) of the ith option on day t and $BAW_i(*)$ is the appropriate call or put formula as a function of the IV.

Before being used in the minimization of (7), the data were checked to make sure that they obeyed the inequality restrictions implied by the no-arbitrage conditions on American options prices: $C \geq F - X$ and $P \geq X - F$, where F is the price of the underlying futures contract. These conditions apply because an American option—which can be exercised at any time—must always be worth at least its value if exercised immediately. Options prices that did not obey these relations were discarded. In addition, the observation was discarded if there was not at least one call and one put price.

THE PROPERTIES OF IMPLIED VOLATILITY

How Well Does IV Predict RV?

Equation (6) says that BS IV is approximately the conditional expectation of RV($\bar{V}_{t,T}$). This relation has two testable implications: IV should be an unbiased predictor of RV; no other forecast should improve the forecast from IV. If IV is an unbiased predictor of RV, one should find that $\{\alpha, \beta_1\} = \{0, 1\}$ in the following regression:

(8) $$\sigma_{RV,t,T} = \alpha + \beta_1 \sigma_{IV,t,T} + \varepsilon_t,$$

where $\sigma_{RV,t,T}$ denotes the RV of the asset return from time t to T and $\sigma_{IV,t,T}$ is IV at t for an option expiring at T.[11] RV is the annualized standard deviation of asset returns from t to T:

(9) $$\sigma_{RV,t,T} = \sqrt{\bar{V}_{t,T}} = \sqrt{\frac{250}{T-t} \sum_{i=t}^{T} \ln(F_i / F_{i-1})},$$

where F_t is the asset price at t and there are 250 business days in the year.

The other commonly investigated hypothesis about IV is that no other forecast improves its forecasts of RV. If IV does subsume other information in this way, it is said to be an "informationally efficient predictor" of volatility. Researchers investigate this issue with variants of the following encompassing regression:

(10) $$\sigma_{RV,t,T} = \alpha + \beta_1 \sigma_{IV,t,T} + \beta_2 \sigma_{FV,t,T} + \varepsilon_t,$$

[10] The results in this paper are almost indistinguishable when done with European option pricing formulas (Black, 1976) or the Barone-Adesi and Whaley correction for American options.

[11] Researchers also estimate (8) with realized and implicit variances, rather than standard deviations. The results from such estimations provide similar inference to those done with variances. Other authors argue that because volatility is significantly skewed, one should estimate (8) with log volatility. Equation (6) shows that use of logs introduces another source of bias into the theoretical relation between RV and IV.

where $\sigma_{FV,t,T}$ is some alternative forecast of volatility from t to T.[12] If one rejects that $\beta_2 = 0$ for some $\sigma_{FV,t,T}$, then one rejects that IV is informationally efficient.

Across many asset classes and sample periods, researchers estimating versions of (8) have found that $\hat{\alpha}$ is positive and $\hat{\beta}_1$ is less than 1 (Canina and Figlewski, 1993; Lamoureux and Lastrapes, 1993; Jorion, 1995; Fleming, 1998; Christensen and Prabhala, 1998; Szakmary et al., 2003). That is, IV is a significantly biased predictor of RV: A given change in IV is associated with a larger change in RV.

Tests of informational efficiency provide more mixed results. Kroner, Kneafsey, and Claessens (1993) concluded that combining time-series information with IV could produce better forecasts than either technique singly. Blair, Poon, and Taylor (2001) discover that historical volatility provides no incremental information to forecasts from VIX IVs.[13] Li (2002) and Martens and Zein (2004) find that intraday data and long-memory models can improve on IV forecasts of RV in currency markets.

It is understandable that tests of informational efficiency provide more varied results than do tests of unbiasedness. Because theory does not restrict what sort of information could be tested against IV, the former tests suffer a data snooping problem. Even if IV is informationally efficient, some other forecasts will improve its predictions in a given sample, purely as a result of sampling variation. These forecasts will not add information to IV in other periods, however.

But some authors have found reasonably strong evidence against the simple informational efficiency hypothesis across assets and classes of forecasts (Neely, 2004a,b). This casts doubt on the data snooping explanation. It seems likely that IV is not informationally efficient by statistical criteria and that the failure of unbiasedness and inefficiency are related.

Several hypotheses have been put forward to explain the conditional bias: errors in IV estimation, sample selection bias, estimation with overlapping observations, and poor measurement of RV. Perhaps the most popular solution to the conditional bias puzzle is the claim that volatility risk is priced. This theory requires some explanation.

The Price of Volatility Risk

To understand the volatility risk problem, consider that there are two sources of uncertainty for an option *writer*—the agent who sells the option—if the volatility of the underlying asset can change over time: the change in the price of the underlying asset and the change in its volatility.[14] An option writer would have to take a position both in the underlying asset (delta hedging) and in another option (vega hedging) to hedge both sources of risk.[15] If the investor only hedges with the underlying asset—not using another option too—then the return to the investor's portfolio is not certain. It depends on changes in volatility. If such volatility fluctuations represent a systematic risk, then investors must be compensated for exposure to them. In this case, the Hull-White result (4) does not apply because there will be risk associated with holding the option and the IV from the BS formula will not approximate the conditional expectation of objective variance as in (6).

The idea that volatility risk might be priced has been discussed for some time: Hull and White (1987) and Heston (1993) consider it. Lamoureux and Lastrapes (1993) argued that the price of volatility risk was likely to be responsible for the bias in IVs options on individual stocks. But most empirical work has assumed that this volatility risk premium is zero, that volatility risk could be hedged or is not priced.

[12] One need not make the econometric forecast orthogonal to IV before using it in (10). The $\hat{\beta}_2$ t-statistic provides the same asymptotic inference as the appropriate F-test for the null that $\beta_2 = 0$. And the F-test is invariant to orthogonalizing the regressors because it is based on the regression R^2.

[13] VIX is a weighted index of IVs calculated from near-the-money, short-term, S&P 100 options. It is designed to correct measurement problems associated with the volatility smile and early exercise.

[14] A more general model would imply additional sources of risk such as discontinuities (jumps) in the underlying asset price or underlying volatility.

[15] Delta and vega denote the partial derivatives of the option price with respect to the underlying asset price and its volatility, respectively.

Is it reasonable to assume that the volatility risk premium is zero? There is no question that volatility is stochastic, options prices depend on volatility, and risk is ubiquitous in financial markets. And if customers desire a net long position in options to hedge against real exposure or to speculate, some agents must hold a net short position in options. Those agents will be exposed to volatility fluctuations. If that risk is priced in the asset pricing model, those agents must be compensated for exposure to that risk. These facts argue that a non-zero price of volatility risk creates IV's bias.

On the other hand, there seems little reason to think that volatility risk itself should be priced. While the volatility of the market portfolio is a priced factor in the intertemporal capital asset pricing model (CAPM) (Merton, 1973a; Campbell, 1993), it is more difficult to see why volatility risk in other markets—e.g., foreign exchange and commodity markets—should be priced. One must appeal to limits-of-arbitrage arguments (Shleifer and Vishny, 1997) to justify a non-zero price of currency volatility risk.

Recently, researchers have paid greater attention to the role of volatility risk in options and equity markets (Poteshman, 2000; Bates, 2000; Benzoni, 2002; Chernov, 2002; Pan, 2002; Bollerslev and Zhou, 2003; and Ang et al., 2003). Poteshman (2000), for example, directly estimated the price of risk function and instantaneous variance from options data, then constructed a measure of IV until expiry from the estimated volatility process to forecast SPX volatility over the same horizon. Benzoni (2002) finds evidence that variance risk is priced in the S&P 500 option market. Using different methods, Chernov (2002) also marshals evidence to support this price of volatility risk thesis. Neely (2004a,b) finds that Chernov's price-of-risk procedures do not explain the bias in foreign exchange and gold markets.

THE IMPLIED VOLATILITY OF SHORT-TERM INTEREST RATES

The IV of options on short-term interest rates illustrates how IV might be applied to understand economic forces. Central banks are particularly concerned with short-term interest rates because most central banks implement monetary policy by targeting those rates.[16] Financial market participants and businesses likewise often carefully follow the actions and announcements of central banks to better understand the future path of short-term interest rates.

Eurodollar Futures Contracts

Interest rate futures are derivative assets whose payoffs depend on interest rates on some date or dates in the future. They enable financial market participants to either hedge their exposure to interest rate fluctuations, or speculate on interest rate changes. One such instrument is the Chicago Mercantile Exchange futures contract for a three-month eurodollar time deposit with a principal amount of $1,000,000. The final settlement price of this contract is 100 less the British Bankers' Association (BBA) three-month eurodollar rate prevailing on the second London business day immediately preceding the third Wednesday of the contract month:

$$F_T = 100 = R_T, \tag{11}$$

where F_T is the final settlement price of the futures contract and R_T is the BBA three-month rate on the contract expiry date. The relation between the three-month eurodollar rate at expiry and the final settlement price ties the futures price at all dates to expectations of this interest rate.

For concreteness, consider what would happen if the First Bank of Des Peres (FBDP) sold a three-month eurodollar futures contract for a quoted price of $97 on June 7, 2004, for a contract expiring on September 13, 2004. Banks might take such short positions to hedge interest rate fluctuations; they borrow short-term and lend long-term and will generally lose (gain) when short-term interest rates rise (fall). The FBDP's

[16] The fact that central banks implement policy by targeting short-term interest rates does not mean that nominal interest rates can be interpreted as measuring the stance of monetary policy. For example, if inflation rises and interest rates remain constant, policy passively becomes more accommodative, all else equal.

short position means that it has effectively agreed to borrow \$1,000,000 for three months, starting on September 13, 2004, at an interest rate of (100 – 97 =) 3 percent.

If the market had expected no change in interest rates through September and risk premia in this market are constant, then realized changes in spot interest rates will translate directly into changes in futures prices.[17] If interest rates unexpectedly rise 45 basis points between June 7, 2004, and September 13, 2004, the FBDP futures prices will fall and the FBDP will have gained by precommitting to borrow at 3 percent. If interest rates unexpectedly decline, however, the FBDP will lose on the futures contract.

How much will the FBDP gain (lose) for each basis-point decrease (increase) in interest rates? With quarterly compounding it will gain 1 basis point of interest for one quarter of a year on \$1,000,000. This translates to \$25 per basis point.

$$\$1,000,000 \frac{0.0001}{4} = \$25. \tag{12}$$

If the BBA three-month eurodollar rate is 3.45 percent on the day of final settlement, the final settlement price of the futures contract will be 100 – 3.45 = 96.55 percent. The FBDP will gain \$25 × 45 = \$1,125 because it shorted the contract at \$97 and the contract price fell to \$96.55 at final settlement.[18] Such a gain would be used to offset losses from the reduced value of its asset portfolio (loans).

Because the final futures price will be determined by the BBA three-month eurodollar rate at final settlement, the futures price can be used to infer the expected future interest rate if there is no risk premium associated with holding the futures contract. Or, if there are stable risk premia associated with holding the contract, one can still measure changes in expected interest rates from changes in futures prices if the risk premia are fairly stable.

Splicing the Futures and Options Data

To examine the behavior of IV on short-term interest rates, we consider settlement data on each three-month eurodollar futures and option contract for the period March 20, 1985, through June 29, 2001. Because exchange-traded futures and options contracts expire on only a few dates a year, one cannot obtain a series of options priced with a fixed expiry horizon for each business day of the year.[19] To obtain as much information as possible, the usual practice in dealing with futures and options data is to “splice” data from different contracts at the beginning of some set of contract expiry months, usually monthly or quarterly. This article uses data from futures and options contracts expiring in March, June, September, and December. For example, settlement prices for the futures contract and the two nearest-the-money call and put options expiring in March 1986 are collected for all trading days in December 1985 and January and February 1986. Then data pertaining to June 1986 contracts are collected from March, April, and May 1986 trading dates. A similar procedure is followed for the September and December contracts. Such a procedure avoids pricing problems near final settlement that result from illiquidity (Johnston, Kracaw, and McConnell, 1991). This method collects data on a total of 4,040 business days, with 8 to 76 business days to option expiry.

Summary Statistics

Table 1 shows the summary statistics on log futures price changes in percentage terms, absolute log futures price changes in annual terms, and IV and RV in annual terms. Futures price changes are very close to mean-zero and have some modest positive autocorrelation. The absolute changes are definitely positively autocorrelated, as one would expect from high-frequency asset price data. IV and RV until expiry have similar mean and autocorrelation properties. But IV is somewhat less volatile than RV, as one would expect if IV predicts RV. The mean of RV is slightly lower

[17] More generally, only unanticipated changes in interest rates will result in changes in futures prices and risk premia will play some role in futures returns.

[18] This example assumes the FBDP holds the position until final settlement.

[19] Additional expiry months were introduced in 1995; previously, there were four expiry months per year.

Table 1

Summary Statistics

	$100 \cdot \ln(F(t)/F(t-1))$	$249 \cdot 100 \cdot \lvert \ln(F(t)/F(t-1)) \rvert$	$\sigma_{IV,t,T}$	$\sigma_{RV,t,T}$
Total observations	4,040	4,040	4,040	4,040
Nobs	3,975	3,975	3,953	4,039
μ	0.003	10.088	0.953	0.769
σ	0.070	14.141	0.458	0.494
Max	1.272	316.645	3.601	3.861
Min	–0.449	0.000	0.251	0.076
ρ_1	0.070	0.213	0.986	0.989
ρ_2	0.023	0.241	0.973	0.977
ρ_3	–0.014	0.226	0.960	0.965
ρ_4	–0.025	0.246	0.948	0.954
ρ_5	–0.007	0.247	0.936	0.942

NOTE: The table contains summary statistics on log futures price changes (percent), annualized absolute log futures price changes, and annualized IV and RV until expiry. The rows show the total number of observations in the sample, the non-missing observations, the mean, the standard deviation, the maximum, the minimum, and the first five autocorrelations. The standard error of the autocorrelations is about $1/\sqrt{T} \approx 0.016$.

than that of IV, indicating that there might be a volatility premium.

Figure 2 clearly illustrates the right skewness in the distribution of IV and changes in IV. Although it is difficult to see in the lower panel of Figure 2, very large positive changes in IV are much more common than very large negative changes in IV. The fact that IV must be positive probably partly explains the right skewness in these distributions.

Eurodollar Rates and the Federal Funds Target Rate

The futures and options data considered here pertain to three-month eurodollar rates. The Fed, however, is more concerned about the federal funds rate, the overnight interbank interest rate used to implement monetary policy, than about other short-term interest rates, such as the eurodollar rate.[20] This is because the federal funds futures prices are often interpreted to provide market expectations of the Fed's near-term policy actions. Short-term interest rates are closely tied together, however, so there might be information about the federal funds rate in three-month eurodollar futures.

Figure 3 shows that, although the three-month eurodollar is much more variable than the federal funds target over a period of a few days, the two series closely tracked each other over periods longer than a few days from March 1985 through June 2001. One can assume that the expected path of the funds rate is closely related to the expected path of the three-month eurodollar rate.[21] And therefore the IV on three-month eurodollars probably tracks the uncertainty about the federal funds target over horizons greater than a few days.

Options on Eurodollar Rates

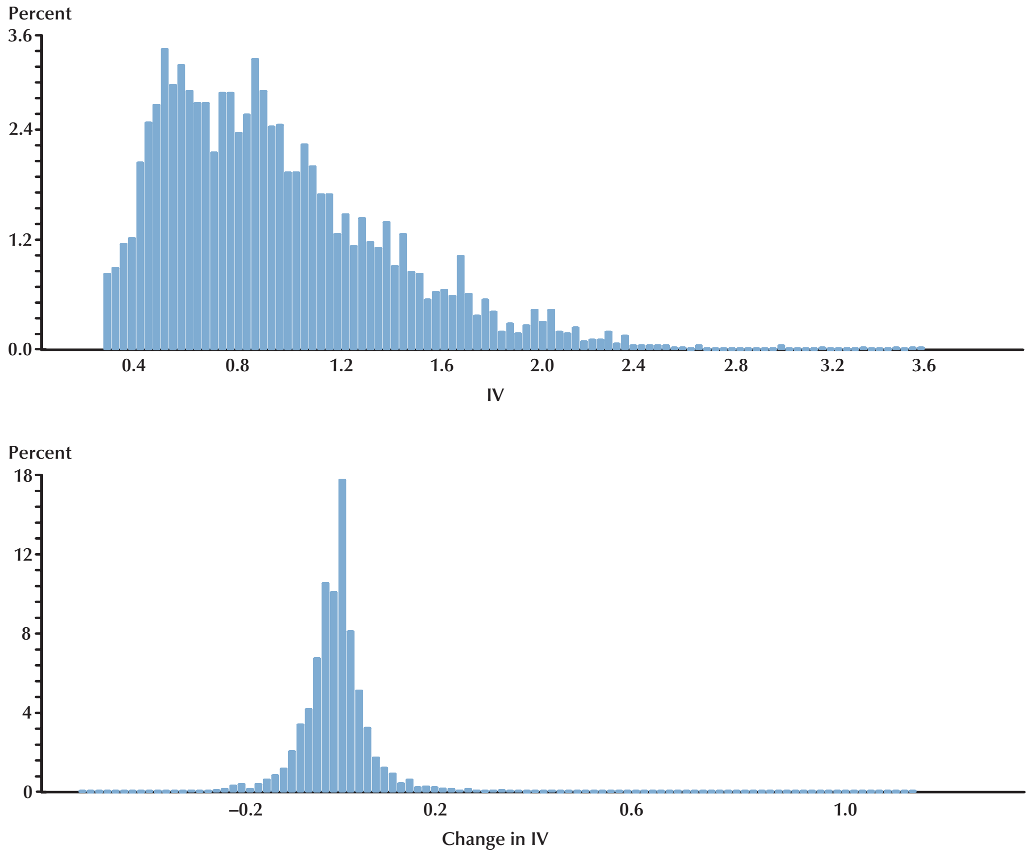
Because option prices depend on the volatility of the underlying asset (among other factors), one can measure the uncertainty associated with expectations of future interest rates from IV from option prices on eurodollar futures contracts. And

[20] Carlson, Melick, and Sahinoz (2003) describe the recently developed options market on federal funds futures contracts.

[21] The payoff to the federal funds futures contract depends on the average federal funds rate over the course of a month, whereas the three-month eurodollar futures contract payoff depends on the BBA quote for the three-month eurodollar rate at one point in time, the expiry of the contract.

Figure 2

The Distributions of Implied Volatility and Changes in Implied Volatility



NOTE: The figure shows the empirical distributions of IV and changes in IV on three-month eurodollar futures prices.

the volatility of interest rates will be very close to the volatility of futures prices because of the linear relation between the two series at final settlement: $100 - F_T = R_T$.

The usual BS measure of IV is a risk-neutral measure, meaning that it assumes that all risk associated with holding the option can be arbitraged away.[22] This is probably not exactly true. And the eurodollars futures prices don't necessarily follow the assumptions of the BS model. In particular, the underlying asset price is probably subject to jumps. Yet Figure 4, which shows the IV and RV until expiry of the three-month eurodollars futures price, appears to show that the BS IV tracks RV fairly well. So, one might think that IV from options on three-month eurodollar rates measures the uncertainty about future interest rates reasonably well.

How Well Does IV Predict RV for Eurodollar Futures?

One can test the unbiasedness hypothesis—that IV is an unbiased predictor of RV—with the predictive regression (8):

$$\sigma_{RV,t,T} = \alpha + \beta_1 \sigma_{IV,t,T} + \varepsilon_t . \quad (8)$$

[22] Boxed insert 2 explains the concept of risk-neutral measures.

BOXED INSERT 2: RISK-NEUTRAL VALUATION

The calculation of the price of the option in Figure 1 did not include any assumptions about the probabilities that the stock price would rise or fall. But the assumptions used to value the stock do imply "risk-neutral probabilities" of the two states of the world. These are the probabilities that equate the expected payoff on the stock with the payoff to a riskless asset that requires the same initial investment. Recall that the stock in the example in Figure 1 was worth \$12 in the first state of the world and \$8 in the second state of the world. If the initial price of the stock is \$10, the risk-neutral probabilities solve the following:

$$p \cdot \$12 + (1-p) \cdot \$8 = \$10e^{0.05}. \tag{b1}$$

This implies that—if prices were unchanged and stocks were valued by risk-neutral investors—the probability that the stock price rises—the probability of state 1—is the following:

$$p = \frac{(10e^{0.05} - 8)}{12 - 8} = 0.6282. \tag{b2}$$

It is important to understand that this risk-neutral probability is not the objective probability that the stock price will rise. It is a synthetic probability that the stock price will rise if actual prices had been determined by risk-neutral agents.

No assumption in this example provides the objective probability that the stock price will rise; neither can one calculate the expected return to the stock. But even without the objective probabilities, one could calculate the option price through the assumption of the absence of arbitrage. It is counter-intuitive but true that the expected return on the stock is not needed to value a call option. One might think that a call option would depend positively on the expected return to the stock. But, because one can value the option through the absence of arbitrage, the expected return to the stock doesn't explicitly appear in the option pricing formula.

And the risk-neutral probabilities can be used to calculate the value of the option (\$$C$) by discounting the value of the (risk-neutral) expected option payoff. Recalling that the option is worth \$2 in the first state of the world, which has a probability of 0.6282 and \$0 in the second state of the world, the option price can be calculated as the discounted risk-neutral expectation of its payoff as follows:

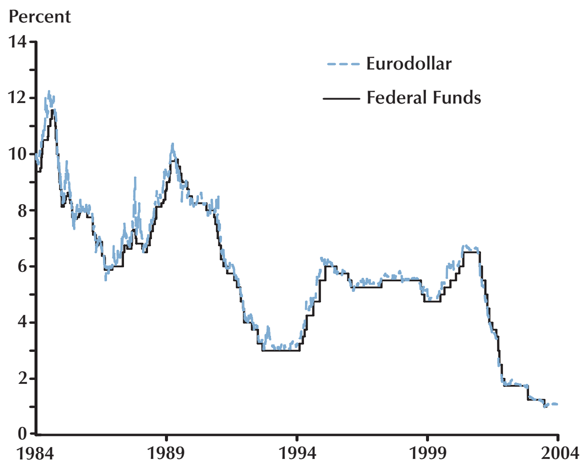
$$C = e^{-0.05}[p \cdot 2 + (1-p) \cdot 0] = e^{-0.05}[0.6282 \cdot 2 + (1-0.6282) \cdot 0] = \$1.1951.$$

This calculation provides the same answer as the no-arbitrage argument used in Figure 1. In some cases, it is easier to derive option pricing formulas from a risk-neutral valuation.

The concept of risk-neutral valuation implies that IV from option prices measures the volatility of the risk-neutral probability measure. To the extent that an asset price's actual stochastic process differs from a risk-neutral process, perhaps because there is a risk-premium in its drift or a volatility risk premium in the option price, the information obtained by inverting option pricing formulas will be misleading. The true distribution of the underlying asset price is often called the *objective* probability measure.

Figure 3

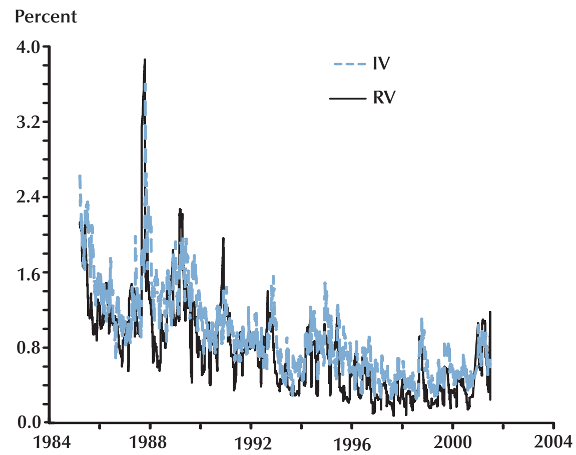
Federal Funds Targets and Three-Month Eurodollar Rates



NOTE: The figure displays federal funds targets and the three-month eurodollar rate from January 1, 1984, to July 25, 2003.

Figure 4

Realized and Implied Volatility on Three-Month Eurodollar Rates



NOTE: The figure displays three-month eurodollar IV and RV from March 20, 1985, through June 29, 2001.

For overlapping horizons, the residuals in (8) will be autocorrelated and, while ordinary least squares (OLS) estimates are still consistent, the autocorrelation must be dealt with in constructing standard errors (Jorion, 1995). Such data sets are described as "telescoping" because correlation between adjacent errors declines linearly and then jumps up at the point at which contracts are spliced.

Table 2 shows the results of estimating (8) with $\sigma_{IV,t,T}$ and $\sigma_{RV,t,T}$ on three-month eurodollar futures. $\hat{\beta}_1$ is statistically significantly less than 1—0.83—indicating that IV is an overly volatile predictor of subsequent RV. This is the usual finding from such regressions: See Canina and Figlewski (1993), Lamoureux and Lastrapes (1993), Jorion (1995), Fleming (1998), Christensen and Prabhala (1998), and Szakmary et al. (2003), for example. As discussed previously, there are many potential explanations for this conditional bias—sample selection, overlapping data, errors in IV—but the most popular story is that stochastic volatility introduces risk to delta hedging, making writing options risky.

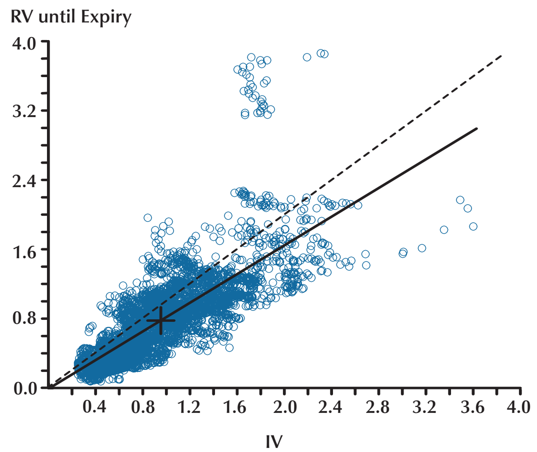
Figure 5 shows a scatterplot of {IV, RV} pairs along with the OLS fitted values from Table 2, a 45-degree line and the mean of IV and RV. If IV were an unbiased predictor of RV, the 45-degree line would be the true relation between them. The fact that the OLS line is flatter than the 45-degree line illustrates that IV is an overly volatile predictor of RV. The cross in Figure 5—which is centered on {mean IV, mean RV}—lies beneath the 45-degree line, illustrating that the mean IV is higher than mean RV.

What Does IV Illustrate About Uncertainty About Future Interest Rates?

Comparing Figure 3 with Figure 4 shows that IV has been declining with the overall level of short-term interest rates, which have been falling with inflation since the early 1980s. One interpretation of the data is that the sharp rise in inflation in the 1970s and the subsequent disinflation of the 1980s created much uncertainty about the level of future interest rates, which has gradually fallen over the past 20 years. The reduction in uncertainty with respect to interest rates probably

Figure 5

Implied Volatility as a Predictor of Realized Volatility



NOTE: The figure shows a scatterplot of {IV, RV} pairs along with the ordinary least-squares fitted values from Table 2 (solid black line), a 45-degree line (short dashes) and the IV and RV (cross). The data are in percentage terms.

Table 2

Predicting Realized Volatility with Implied Volatility

$\hat{\alpha}$	–0.017
(s.e.)	0.052
$\hat{\beta}_1$	0.834
(s.e.)	0.064
Wald	40.814
Wald PV	0.000
Observations	3,952
R^2	0.599

NOTE: The table shows the results of predicting three-month eurodollar RV with IV, as in (8). The rows show $\hat{\alpha}$, its robust standard error, $\hat{\beta}_1$, its robust standard error, the Wald test statistic for the null that $\{\alpha, \beta\} = \{0, 1\}$, the Wald test p-value, the number of observations, and the R^2 of the regression.

stems from both a reduction in the level of interest rates and greater certainty about both monetary policy and the level of real economic activity.

A close look at Figure 4 also hints that there might be some seasonal pattern in IV, associated with the expiry of contracts. Indeed, long-horizon IVs tend to be larger than short-horizon IVs (which, for brevity, are not shown). As IV is scaled to be interpretable as an annual measure, comparable at any horizon, this is a bit of a mystery. It might simply be an artifact of the simplifying assumptions of the BS model.

What Sort of News Is Coincident with Changes in IV?

Events of obvious economic importance and large changes in the futures price, itself, often accompany the largest changes in IV. To examine news events around large changes, the *Wall Street Journal* business section was searched for news on the dates of large changes and on the days immediately following those changes—from March 20, 1985, through June 29, 2001. Table 3 shows some of the largest changes in IV during the sample and the event that might have precipitated it.

The largest change in IV, by far, is a rise of 1.2 percentage points on October 20, 1987, coinciding with the stock market crash of 1987, when the S&P 500 lost 22 percent of its value in one day. Four more of the top 20 changes (including the second largest) happened in the six weeks following the crash and one happened eight weeks before the crash, on August 27, 1987. The large changes in the IV of three-month eurodollar interest rates reflected uncertainty about future interest rates prior to the crash. A change in Federal Reserve Chairmen might have fueled the apparent uncertainty about the economy and the stance of monetary policy. Alan Greenspan took office as Chairman of the Board of Governors of the Federal Reserve on August 11, 1987.

The third largest change, a 0.44-percentage-point increase, occurred on November 28, 1990. It coincided with reports that President George H.W. Bush would go to Congress to ask for endorsement of plans to use military force to evict Iraqi forces from Kuwait. The possibility of war in such an economically important area of the world clearly spooked financial markets.

Table 3

News Events Coincident with Large Changes in Three-Month Eurodollar IV

Rank	Δ in IV	Date	Δ in federal funds target?	Relevant financial news
1	1.182	10/20/87	No	Stock market crash of 1987: S&P 500 declined 22 percent in one day.
2	–0.526	11/12/87	No	Decline in U.S. trade deficit.
3	0.438	11/28/90	No	Gulf War fears: Bush going to Congress to ask for authority to evict Iraq from Kuwait.
4	0.411	8/27/98	No	Russian debt crisis: Yeltsin may resign, along with an indefinite suspension of ruble trading and fear Russia may return to Soviet-style economics.
5	–0.375	1/15/88	No	The sharp narrowing of the trade deficit triggered market rallies.
6	0.353	12/2/96	No	Retailers reported stronger-than-expected sales over Thanksgiving.
7	0.339	10/15/87	No	Stocks and bonds slid further as Treasury Secretary Baker tried to calm the markets, saying the rise in interest rates isn't justified.
8	0.332	9/3/85	No	The farm credit system is seeking a federal bailout of its $74 billion loan portfolio…As much as 15 percent of its loans are uncollectible.
9	0.330	11/27/87	No	Inflation worries remain despite the stock crash, due to higher commodity prices and the weak dollar.
10	–0.321	10/29/87	Yes	Post-stock market crash reduction in the federal funds target.
11	0.315	6/7/85	No	Bond prices declined for the first time in a week, as investors awaited a report today on May employment…The Fed reported a surge in the money supply, leaving it well above the target range.
12	–0.302	10/30/87	No	Stocks and bonds reversed course after an early slide, helped by G-7 interest-rate drops.
13	–0.301	8/16/94	Yes	FOMC meeting: The Fed boosted the funds rate 50 basis points, sending a clear inflation-fighting message.
14	–0.298	7/11/86	Yes	The Fed's discount-rate cut prompted major banks to lower their prime rates.
15	–0.285	12/2/91	No	Under strong pressure to resuscitate the economy, President George H.W. Bush promised not to do "anything dumb" to stimulate the economy.
16	0.279	4/20/89	No	Financial markets were roiled by a surprise half-point boost in West German interest rates. The tightening was quickly matched by other central banks.
17	0.278	8/27/91	No	Federal funds target rate was increased on August 6 and September 13, 1991.
18	0.275	8/31/89	No	Federal funds target rate was increased on August 20 and October 18, 1989.
19	0.275	8/27/87	Yes	Federal funds target rate raised by 12.5 basis points.
20	0.266	6/2/86	No	Bond prices tumbled amid concern the economy will speed up, renewing inflation.

NOTE: The table contains the largest changes in IV (in percentage points, in descending order) and the news (as reported in the *Wall Street Journal*) that was associated with those changes. The sample includes changes from March 20, 1985, through June 29, 2001.

Another large increase, of 0.41 percentage points, occurred on August 28, 1998. This rise was coincident with the Russian debt crisis, rumors that President Yeltsin had resigned, and the possibility of a reversal of Russian political and economic reforms. The Russian debt crisis had potentially serious implications for international investors. Neely (2004c) discusses the episode and its potential effect on U.S. financial markets.

Several of the 20 largest changes in three-month eurodollar IV were also associated with large changes in the futures price. It is likely that these changes in the futures price were unanticipated because large, anticipated changes in futures prices provide profit-making opportunities. Additionally, anticipated changes are unlikely to cause a substantial revision to IV. Four of the 20 largest changes in IV were also associated with presumably unanticipated changes in the federal funds target rate. It seems that unanticipated monetary policy can be an important determinant of uncertainty about future interest rates.

Finally, one might note that the large IV changes shown in Table 3 refute the BS assumptions of a constant or even continuous volatility process. As such, they might be partly responsible for delta hedging errors, which require a risk premium that causes IV to be a conditionally biased estimate of RV.

CONCLUSION

This article has explained the concept of IV and applied it to measure uncertainty about three-month eurodollar rates. The IVs associated with three-month eurodollars can be interpreted to reflect uncertainty about the Federal Reserve's primary monetary policy instrument, the federal funds target rate.

As with IV in most financial markets, the IV of the three-month eurodollar rate has been an overly volatile predictor of RV. IV on the three-month eurodollar rates has been declining since 1985, as inflation and interest rates have fallen and the Fed has gained credibility with financial markets. The largest changes in IV were coincident with important economic events such as the stock-market crash of 1987, fears of war in the Persian Gulf, and the Russian debt crisis. Most of the rest of the largest changes in IV have similarly been associated with important news about the real economy or the stock market or revisions to expected monetary policy.

REFERENCES

Ang, Andrew; Hodrick, Robert J.; Xing, Yuhang and Zhang, Xiaoyan. "The Cross-Section of Volatility and Expected Returns." Unpublished manuscript, Columbia University, 2003.

Barone-Adesi, Giovanni and Whaley, Robert E. "Efficient Analytic Approximation of American Option Values." *Journal of Finance*, 1987, *42*, pp. 301-20.

Bates, David S. "Testing Option Pricing Models," in G.S. Maddala and C.R. Rao, eds., *Statistical Methods in Finance/Handbook of Statistics*. Volume 14. Amsterdam: Elsevier Publishing, 1996.

Bates, David S. "Post-'87 Crash Fears in the S&P 500 Futures Option Market." *Journal of Econometrics*, 2000, *94*, pp. 181-238.

Bates, David S. "Empirical Option Pricing: A Retrospection." *Journal of Econometrics*, 2003, *116*, pp. 387-404.

Beckers, Stan. "Standard Deviations Implied in Options Prices as Predictors of Futures Stock Price Variability." *Journal of Banking and Finance*, 1981, *5*, pp. 363-81.

Benzoni, Luca. "Pricing Options Under Stochastic Volatility: An Empirical Investigation." Unpublished manuscript, Carlson School of Management, 2002.

Black, Fischer. "The Pricing of Commodity Contracts." *Journal of Financial Economics*, 1976, *3*, pp. 167-79.

Black, Fischer and Scholes, Myron. "The Valuation of Option Contracts and a Test of Market Efficiency." *Journal of Finance*, 1972, *27*, pp. 399-417.

Blair, Bevan J.; Poon, Ser-Huang and Taylor, Stephen J. "Forecasting S&P 100 Volatility: The Incremental Information Content of Implied Volatilities and

High-Frequency Index Returns." *Journal of Econometrics*, 2001, *105*, pp. 5-26.

Bollerslev, Tim and Zhou, Hao. "Volatility Puzzles: A Unified Framework for Gauging Return-Volatility Regressions." Finance and Economics Discussion Series 2003-40, Board of Governors of the Federal Reserve System, 2003.

Boyle, Phelim and Boyle, Feidhlim. *Derivatives: The Tools that Changed Finance*. London: Risk Books, 2001.

Campbell, John Y. "Intertemporal Asset Pricing without Consumption Data." *American Economic Review*, 1993, *83*, pp. 487-512.

Canina, Linda and Figlewski, Stephen. "The Informational Content of Implied Volatility." *Review of Financial Studies*, 1993, *6*, pp. 659-81.

Carlson, John B.; Melick, William R. and Sahinoz Erkin Y. "An Option for Anticipating Fed Action." Federal Reserve Bank of Cleveland *Economic Commentary*, September 1, 2003, pp. 1-4.

Chernov, Mikhail. "On the Role of Volatility Risk Premia in Implied Volatilities Based Forecasting Regressions." Unpublished manuscript, Columbia University, 2002.

Christensen, B.J. and Prabhala, N.R. "The Relation Between Implied and Realized Volatility." *Journal of Financial Economics*, 1998, *50*, pp. 125-50.

Dubofsky, David A. and Miller Jr, Thomas M. *Derivatives: Valuation and Risk Management*. New York: Oxford University Press, 2003.

Fleming, Jeff. "The Quality of Market Volatility Forecasts Implied by S&P 100 Index Option Prices." *Journal of Empirical Finance*, 1998, *5*, pp. 317-45.

Heston, Steven L. "A Closed-Form Solution for Options with Stochastic Volatility with Applications to Bond and Currency Options." *Review of Financial Studies*, 1993, *6*, pp. 327-43.

Hull, John C. *Options, Futures, and Other Derivatives*. 5th edition. Upper Saddle River, NJ: Prentice Hall, 2002.

Hull, John C. and White, Alan. "The Pricing of Options on Assets with Stochastic Volatilities." *Journal of Finance*, 1987, *42*, pp. 281-300.

Jarrow, Robert and Turnbull, Stuart. *Derivative Securities*. Cincinnati, OH: South-Western College Publishing, 2000.

Johnston, E.; Kracaw, W. and McConnell, J. "Day-of-the-Week Effects in Financial Futures: An Analysis of GNMA, T-Bond, T-Note, and T-Bill Contracts." *Journal of Financial and Quantitative Analysis*, 1991, *26*, pp. 23-44.

Jorion, Philippe. "Predicting Volatility in the Foreign Exchange Market." *Journal of Finance*, 1995, *50*, pp. 507-28.

Kroner, Kenneth F.; Kneafsey, Kevin P. and Claessens, Stijn. "Forecasting Volatility in Commodity Markets." Working Paper 93-3, University of Arizona, 1993.

Krueger, Joel T. and Kuttner, Kenneth N. "The Fed Funds Futures Rate as a Predictor of Federal Reserve Policy," Working Paper WP-95-4, Federal Reserve Bank of Chicago, March 1995.

Lamoureux, Christopher G. and Lastrapes, William D. "Forecasting Stock-Return Variance: Toward an Understanding of Stochastic Implied Volatilities." *Review of Financial Studies*, 1993, *6*, pp. 293-326.

Latane, Henry A. and Rendleman, Richard J. Jr. "Standard Deviations of Stock Price Ratios Implied in Option Prices." *Journal of Finance*, 1976, *31*, pp. 369-81.

Li, Kai. "Long-Memory versus Option-Implied Volatility Prediction." *Journal of Derivatives*, 2002, *9*, pp. 9-25.

Martens, Martin and Zein, Jason. "Predicting Financial Volatility: High-Frequency Time-Series Forecasts vis-à-vis Implied Volatility." *Journal of Futures Markets*, 2004, *24*(11), pp. 1005-28.

Merton, Robert C. "An Intertemporal Capital Asset Pricing Model." *Econometrica*, 1973a, *41*, pp. 867-87.

Merton, Robert C. "Theory of Rational Option Pricing." *Bell Journal of Economics*, 1973b, *4*(1), pp. 141-83.

Neely, Christopher J. "Realignments of Target Zone Exchange Rate Systems: What Do We Know?" Federal Reserve Bank of St. Louis *Review*, September/October 1994, *76*(5), pp. 23-34.

Neely, Christopher J. "Forecasting Foreign Exchange Volatility: Why Is Implied Volatility Biased and Inefficient? And Does It Matter?" Working Paper 2002-017D, Federal Reserve Bank of St. Louis, 2004a.

Neely, Christopher J. "Implied Volatility from Options on Gold Futures: Do Econometric Forecasts Add Value or Simply Paint the Lilly?" Working Paper 2003-018C, Federal Reserve Bank of St. Louis, 2004b.

Neely, Christopher J. "The Federal Reserve Responds to Crises: September 11th Was Not the First." Federal Reserve Bank of St. Louis *Review*, March/April 2004c, *86*(2), pp. 27-42.

Pakko, Michael R. and Wheelock, David. "Monetary Policy and Financial Market Expectations: What Did They Know and When Did They Know It?" Federal Reserve Bank of St. Louis *Review*, July/August 1996, *78*(4), pp. 19-32.

Pan, Jun. "The Jump-Risk Premia Implicit in Options: Evidence from an Integrated Time-Series Study." *Journal of Financial Economics*, 2002, *63*, pp. 3-50.

Poteshman, Allen M. "Forecasting Future Volatility from Option Prices." Unpublished manuscript, Department of Finance, University of Illinois at Urbana-Champaign, 2000.

Romano, Marc and Touzi, Nizar. "Contingent Claims and Market Completeness in a Stochastic Volatility Model." *Mathematical Finance*, 1997, *7*, pp. 399-410.

Rose, Andrew K. and Svensson, Lars E.O. "Expected and Predicted Realignments: The FF/DM Exchange Rate During the EMS." International Finance Discussion Paper Number 395, Board of Governors of the Federal Reserve System, April 1991.

Rose, Andrew K. and Svensson, Lars E.O. "European Exchange Rate Credibility Before the Fall." NBER Working Paper No. 4495, National Bureau of Economic Research, October 1993.

Shleifer, Andrei, and Vishny, Robert W. "The Limits of Arbitrage." *Journal of Finance*, 1997, *54*, pp. 35-55.

Svensson, Lars E.O. "The Simplest Test of Target Zone Credibility." *IMF Staff Papers*, September 1991, pp. 655-65.

Szakmary, Andrew; Ors, Evren; Kim, Jin Kyoung and Davidson, Wallace N. III. "The Predictive Power of Implied Volatility: Evidence from 35 Futures Markets." *Journal of Banking and Finance*, 2003, *27*, pp. 2151-75.

Wilmott, Paul; Howison, Sam and Dewynne, Jeff. *The Mathematics of Financial Derivatives: A Student Introduction.* Cambridge: Cambridge University Press, 1995.

Wilmott, Paul. *Paul Wilmott on Quantitative Finance*. Chichester, UK: John Wiley & Sons, 2000.

GLOSSARY

A ***European option*** is an asset that confers the right, but not the obligation, to buy or sell an underlying asset for a given price, called a ***strike price***, at the ***expiry*** of the option.

An ***American option*** can be exercised on or before the expiry date.[1]

Call options confer the right to buy the underlying asset; ***put options*** confer the right to sell the underlying asset.

If the underlying asset price is greater (less) than the strike price, a call (put) option is said to be ***in the money***. If the underlying asset price is less (greater) than the strike price, the call (put) option is ***out of the money***. When the underlying asset price is near (at) the strike price, the option is ***near (at) the money***.

The firm or individual who sells an option is said to ***write*** the option.

The price of an option is often known as the option ***premium***.

---

[1] The terms European and American no longer have any geographic meaning when referring to options. That is, both types of options are traded worldwide.

Federal Reserve Bank of St. Louis
P.O. Box 442
St. Louis, MO 63166-0442