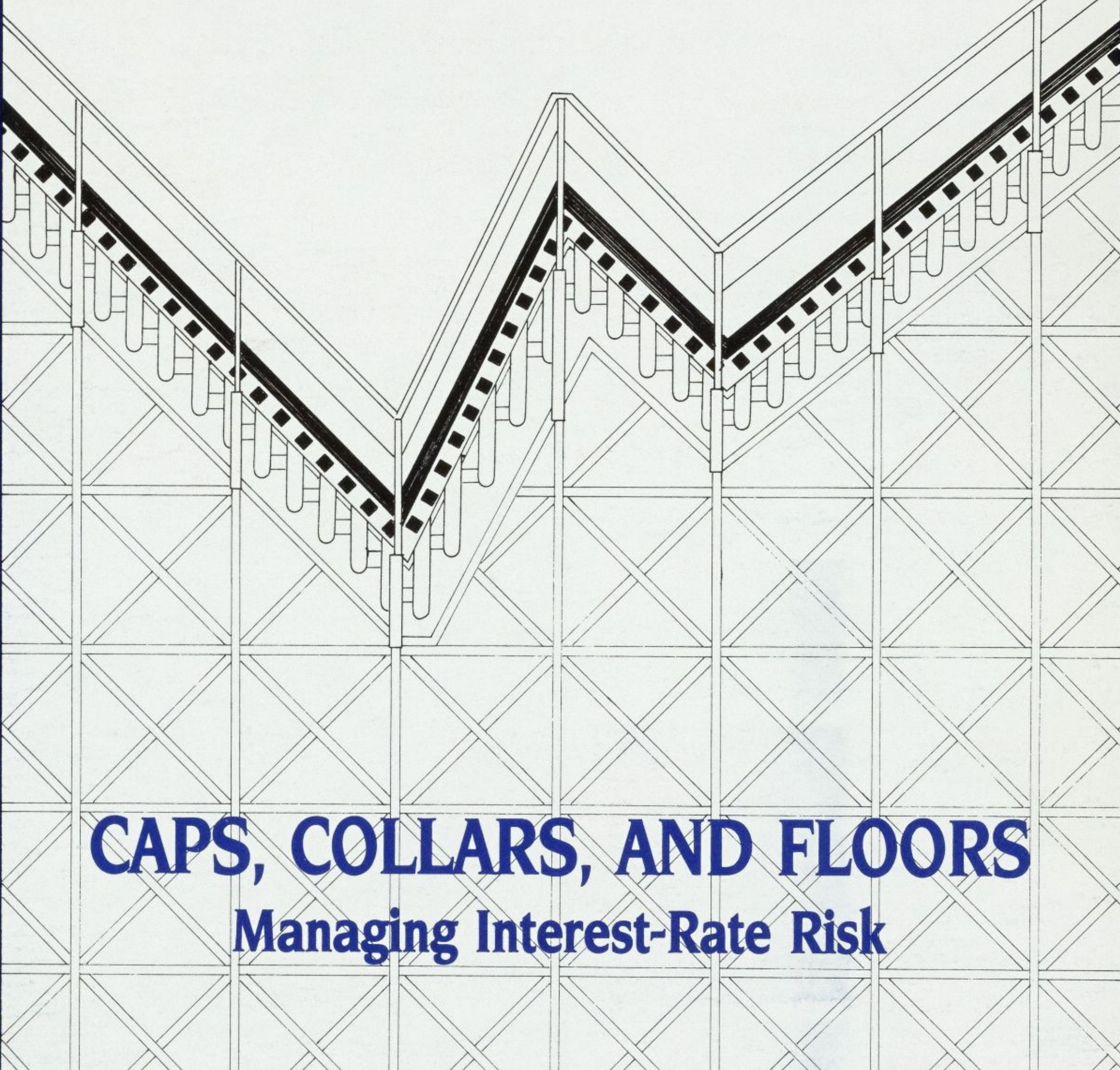
Economic Review

FEDERAL RESERVE BANK OF ATLANTA NOVEMBER/DECEMBER 1989

Financial Asset Pricing

Foreign Direct Investment Update



CAPS, COLLARS, AND FLOORS

Managing Interest-Rate Risk

FEDERAL RESERVE BANK OF ATLANTA
75 YEARS 1914-1989

Economic Review

The **Economic Review** seeks to inform the public about Federal Reserve policies and the economic environment and, in particular, to narrow the gap between specialists and concerned laypersons.

Views expressed in the **Economic Review** are not necessarily those of this Bank or of the Federal Reserve System.

Free subscriptions and limited additional copies are available from the Public Information Department, Federal Reserve Bank of Atlanta, 104 Marietta Street, N.W., Atlanta, Georgia 30303-2713 (404/521-8788).

Change-of-address notices and subscription cancellations should be sent directly to the Public Information Department. Please include the current mailing label as well as any new information.

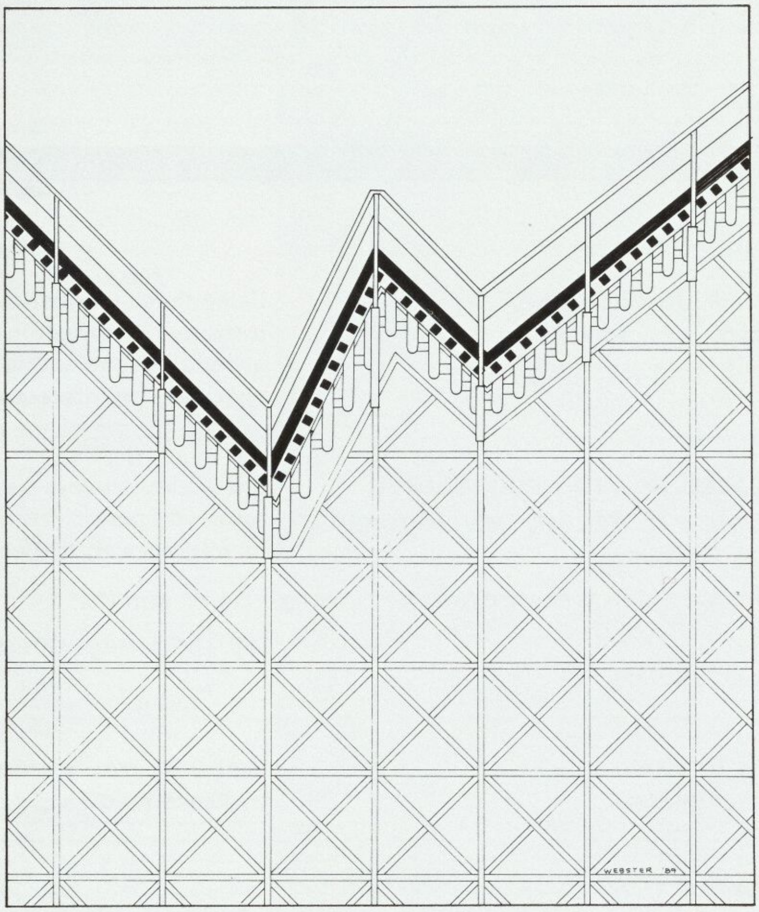
ISSN 0732-1813

VOLUME LXXIV, NO. 6, NOVEMBER/DECEMBER 1989, ECONOMIC REVIEW

FEDERAL RESERVE BANK OF ATLANTA

Interest-Rate Caps, Collars, and Floors

Peter A. Abken



As some of the newest interest-rate risk management instruments, caps, collars, and floors are the subject of increasing attention among both investors and analysts. This article explains how such instruments are constructed, discusses their credit risks, and presents a new approach for valuing caps, collars, and floors subject to default risk.

Since the late 1970s interest rates on all types of fixed-income securities have become more volatile, spawning a variety of methods to mitigate the costs associated with interest-rate fluctuations. Managing interest-rate risk has become big business and an exceedingly complicated activity. One facet of this type of risk management involves buying and selling "derivative" assets, which can be used to offset or hedge changes in asset or liability values caused by interest-rate movements. As its name implies, the value of a derivative asset depends on the value of another asset or assets.

Two types of derivative assets widely discussed in the financial press and in previous *Economic Review* articles are options and futures contracts.[1] Another derivative asset that has become extremely popular is the interest-rate swap.[2] This article examines a group of instruments known as interest-rate caps, collars, and floors, which are medium- to long-term agreements that have proven to be highly useful for hedging against interest-rate uncertainties. In this regard, caps, collars, and floors can be thought of as insurance policies against adverse movements in interest rates.

Like interest-rate swaps, to which these instruments are closely related, caps, collars, and floors are designed to hedge cash flows over time rather than on a single date. The discussion below will show how caps, collars, and floors are related to each other, as well as how they may be constructed from the most basic derivative asset, the option. The article also shows the ways in which caps, collars, and floors are created in practice, along with the different kinds of intermediaries involved in the cap market.[3] The rationale for hedging is reviewed, as are examples of how caps, collars, and floors are used by different financial institutions. The last section of the article considers the credit risk associated with buying caps, collars, or floors and presents a new approach for determining the expected cost of default on these instruments.

The author is an economist in the financial section of the Atlanta Fed's Research Department. He thanks Igor A. Lamser of Noonan, Astley, and Pearce, Inc., for helpful discussions about the cap market and for providing data on cap rates.

What Is an Interest-Rate Cap?

An interest-rate cap, sometimes called a ceiling, is a financial instrument that effectively places a maximum amount on the interest payment made on floating-rate debt. Many businesses borrow funds through loans or bonds on which the periodic interest payment varies according to a prespecified short-term interest rate. The most widely used rate in both the caps and swaps markets is the London Interbank Offered Rate (LIBOR), which is the rate offered on Eurodollar deposits of one international bank held at another.[4] A typical example of floating-rate borrowing might be a firm taking out a $20 million bank loan on which the interest would be paid every three months at 50 basis points (hundredths of a percent) over LIBOR prevailing at each payment date. Other short-term rates that are used in conjunction with caps include commercial bank certificate of deposit (CD) rates, the prime interest rate, Treasury bill rates, commercial paper rates, and certain tax-exempt interest rates.

Data on the size of the cap market are sketchy. The International Swap Dealers Association (ISDA) conducted a survey of its members in March 1989, and 44 of the association's 97 members responded. Almost 90 percent of the respondents reported participating in the markets for caps, collars, floors, and options on swaps. As of year-end 1988, these members alone held 7,521 caps, collars, and floors, with a total notional principal of $290 billion. The volume conducted through 1988 was reported as having notional principal of $172 billion. These figures inflate the size of the market considerably because they are not adjusted for transactions among the dealers themselves, such as the purchase or sale of caps or floors to hedge existing positions in these instruments. On the other hand, the survey did not cover the entire market. Nonetheless, the figures probably still greatly overstate the size of the market, net of interdealer transactions or positions.[5] The interest-rate swaps market is vastly larger at over $1 trillion.

Most studies of caps concern agreements offered by commercial or investment banks to borrowers seeking interest-rate protection. These instruments are often tailored to a client's needs, and, particularly in the case of caps, may be marketable or negotiable. Caps, collars, and floors can also be manufactured out of basic derivative assets: options or futures contracts, or a combination of the two. The following discussion will define caps, collars, and floors in terms of option contracts, which are the simplest type of derivative asset.

Call and Put Options. An option is a financial contract with a fixed expiration date that offers either a positive return (payoff) or nothing at maturity, depending on the value of the asset underlying the option. At expiration, a call option gives the purchaser the right, but not the obligation, to buy a fixed number of units of the underlying asset if that asset's price exceeds a level specified in the option contract. The seller or "writer" of a call has the obligation to sell the underlying asset at the specified exercise or strike price if the call expires "in the money." The payoff on a call need not actually involve delivery of the underlying asset to the call buyer but rather can be settled by a cash payment. The caps market, for example, uses cash settlement. If the asset price finishes below the exercise price, the call is said to expire "out of the money."

Put options are analogous to calls. In this case, though, the purchaser has the right to sell, rather than buy, a fixed number of units of the underlying asset if the asset price is below the exercise price. The options discussed in this article will all be "European" options, which can only be exercised on the expiration date, as opposed to "American" options, which can be exercised any time before or at expiration. As will be seen, caps, floors, and collars are European-style option-based instruments, and the European interest-rate call option is the basic building block for the interest-rate cap.

Options on debt instruments can be confusing if it is unclear just what the option "price" represents. For debt instruments, the strike price is referred to as the strike level, reflecting an interest rate. Recall that the price of a debt instrument, such as a Treasury bill or CD, moves inversely with its corresponding interest rate; as the interest rate of a Treasury bill rises, its price falls. Thus, a call on a Treasury bill rate is effectively a put on its price. (To keep the exposition clear, all discussion will be in terms of options on interest rates. The strike price will be referred to as the strike level.) A call with a strike level of 8 percent (on an annual basis) on some notional amount of principal is effectively a cap on a floating-rate loan payment coinciding with the expiration of this option. (The notional amount of principal is a sum used as the basis for the option payoff computation. Cap, collar, and floor agreements do not involve any exchange of principal.)

Assume the call's payment date, known as the reset date, falls semiannually. If the interest rate is less than 8 percent on the reset date, the call expires worthless. If the interest rate exceeds 8 percent, the call pays off the difference between the actual interest rate and the strike level times the notional principal, in turn multiplied by the fraction of a year that has elapsed since purchase of the option. For example, if the actual rate of interest six months later were 10 percent and if the notional principal were \$1 million, the payment received from the call writer would be 2 percent (the 10 percent actual rate minus the 8 percent strike level) x \$1,000,000 x 180/360 = \$10,000.

"[C]aps, floors, and collars are European-style option-based instruments, and the European interest-rate call option is the basic building block for the interest-rate cap."

A put option on an interest payment works in a similar way and is the foundation for the interest-rate floor. The holder of a floating-rate loan could protect against a loss in interest income from the loan by buying an interest-rate put. A fall in the interest rate below the strike level of the put would result in a payoff from the option, offsetting the interest income lost because of a lower interest payment on the loan.

An option writer is basically an insurer who receives a premium payment from the option buyer when an option is created (sold). In fact, the option price is alternatively called the option premium. The same party can simultaneously write and buy options, thus creating an interest-rate collar. Before exploring this strategy further, option pricing must be reviewed briefly.

Option Pricing. An option's price before expiration depends on several variables, including the value of the underlying asset on which the option is written, the risk-free rate of interest (usually a Treasury bill that matures at the same time as the option), the time remaining before expiration, the strike price or level, and the volatility of the underlying asset price.[6] For later reference, readers should know how an option price changes in response to a change in an underlying variable, all other variables remaining constant. A call price rises (falls) when the underlying asset price, volatility, or time to expiration increases (decreases). It falls (rises) with an increase (decrease) in the exercise price. A put price rises (falls) with an increase (decrease) in the strike price or volatility. It falls (rises) with an increase (decrease) in the underlying asset price or interest rate. Unlike a call price, a put price is not unambiguously affected by an increase in the time to expiration, but the put price depends at any time on how far in or out of the money the put is.[7]

"A cap can . . . be perceived as a series of interest-rate call options for successively more distant reset dates; a floor is a similarly constructed series of put options."

For an interest-rate call option, the higher the strike level compared to the current interest rate, the lower the option value. Choosing a high strike level (out-of-the-money) call is less expensive than buying an at-the-money or in-the-money call. Similarly, a low strike level (out-of-the-money) put is cheaper than one with a higher strike level.

This relationship between an option's strike level and its price (the amount the option is out of the money) is analogous to a large deductible on an insurance policy. Such a policy is less likely to pay off and is therefore less expensive. Likewise, the cost of interest-rate "insurance" can be reduced by taking a large deductible—that is, buying an out-of-the-money option—and thereby protecting only against large, adverse interest-rate movements.

Creating an interest-rate collar is another method for reducing the cost of interest-rate insurance. The call-option premium for an interest-rate cap may be partially or completely offset by selling a put option that sets an interest-rate floor. For a floating-rate debt holder, the effect of this dual purchase is to protect against rate movements above the cap level while simultaneously giving up potential interest savings if the rate drops below the floor level.

If the cap and floor levels of a collar are narrowed to the extent that they coincide at the current floating interest rate—that is, both put and call options are at the money—the resulting collar is so tight that it is similar to a forward contract on an interest rate, which is a derivative asset that locks in the current forward rate. When the contract expires, the change in the contract's value that has occurred since the inception of the contract exactly offsets the change in the interest payment due. A rise in the floating-rate payment is matched by an equal gain in the interest paid to the contract holder; a fall in the floating-rate payment is balanced by an equal loss on the forward contract. In effect, a forward contract converts a floating-rate payment to a fixed-rate payment.

The discussion thus far has been about a single payment, yet, as mentioned earlier, actual cap, collar, or floor agreements are designed to hedge a series of cash flows, not just one. A cap can thus be perceived as a series of interest-rate call options for successively more distant reset dates; a floor is a similarly constructed series of put options. Assume that an interest payment on floating-rate debt falls due in three months, at the next reset date. If the interest rate on the reset date exceeds the strike level, the cap writer would make a payment to the cap buyer on a date to coincide with the cap buyer's own payment date on the underlying floating-rate debt.

A collar that consists of a series of at-the-money call and put options is equivalent to an interest-rate swap. Buying the cap and selling the floor transforms floating-rate debt to fixed-rate debt, whereas selling the cap and buying the floor switches fixed-rate debt into floating-rate debt. A swap that is constructed out of cap and floor agreements is called a *synthetic swap*. Caps brokers and dealers will sometimes determine rates on floors by deriving the rate from swap and cap rates, which come from instruments that are more actively traded than floors and therefore more accurately reflect current market values.

In practice, swaps are not usually put together from cap and floor agreements. Caps and floors are more readily tradable than swaps because credit risk is one-sided; swaps carry a credit risk that is two-sided in nature. Matching buyers and sellers for swaps is therefore more involved than for caps or floors.[8]

Examples of some caps, collars, and floors should help the reader understand their operation. As the foregoing single-payment-date discussion illustrates, creating these instruments amounts to an exercise in option pricing. One widely used option-pricing model, known as the Black futures option model, is used in the following examples.[9] Robert Tompkins (1989) explains caps pricing in terms of Black's model, and the examples that follow are loosely patterned on Tompkins' approach.

The chief virtue of the Black model is its simplicity and ease of use, even though it has a serious internal inconsistency when used to value debt options: the assumption that the short-term interest rate (that is, the Treasury bill rate) is constant. Options on short-term interest rates have value, though, only if those rates are less than perfectly predictable. In the last section of this paper, a more complex model that does not suffer from this shortcoming is used to price options.[10]

Eurodollar Futures and Forward LIBOR. In order to give realistic yet simple examples of caps, collars, and floors, this article assumes that the reset dates coincide with the expiration dates of Eurodollar futures contracts, which are traded at the Chicago Mercantile Exchange (CME) and the London International Financial Futures Exchange (LIFFE). Purchase of a Eurodollar futures contract locks in the interest payment on a $1 million three-month time deposit to be made upon expiration of the futures contract. The interest rate on the deposit is three-month LIBOR. On the other hand, the seller of a Eurodollar futures contract is obligated to pay the specified LIBOR-based interest payment at expiration.[11]

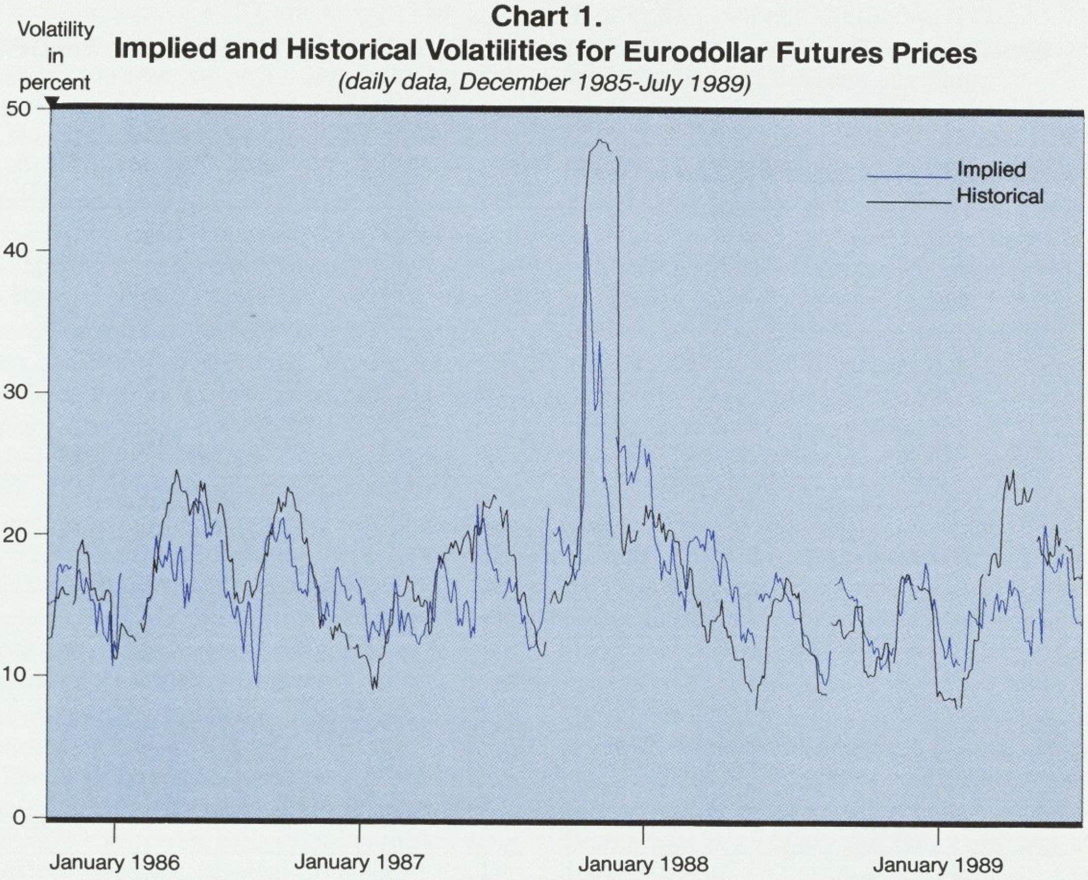
Eurodollar futures expire in a quarterly cycle two London business days prior to the third Wednesday of March, June, September, and December. The Chicago Mercantile Exchange currently offers contract expiration months extending four years, with only March and September contracts for the fourth year.[12] The interest rate implied by a Eurodollar futures price may be regarded as a forward interest rate, that is, the three-month LIBOR expected by the market to prevail at the expiration date for each contract.[13]

The Black model uses the futures price for a particular contract expiration month as an input to determine the value of a European call and put option on that contract. In the case of Eurodollar futures contracts, the add-on yield (100 minus the futures price) is plugged into Black's formula. Another crucial variable is the volatility, which is either estimated from the historical volatility of the Eurodollar futures yield or obtained as an implied volatility from traded Eurodollar futures options.[14] Chart 1 shows the recent behavior of both of these volatility measures. Again, higher volatility results in higher-cost call and put options and hence more expensive caps and floors.

Table 1 gives two-year cap, floor, and collar prices on three-month LIBOR for two arbitrarily chosen dates, June 19, 1989, and December 14, 1987, that give reset dates which coincide with Eurodollar futures expiration dates. The first date illustrates pricing during a relatively low volatility period when the term structure of LIBOR rates, as given by the "strip" of prices on successively more distant contracts, was just about flat. The market was predicting virtually no change in short-term interest rates over this two-year horizon. In panel A of Table 1, the contract expiration months are given along with the forward rates or add-on yields for each futures contract. The row labeled *time to expiration* shows the number of days from the creation of the cap, floor, or collar to the expiration date for each contract. Another input into Black's formula, the risk-free rate, is taken to be the Treasury bill or zero-coupon bond yield for which the

Chart 1.
Implied and Historical Volatilities for Eurodollar Futures Prices

(daily data, December 1985-July 1989)



Higher volatility, such as that exhibited around the time of the October 1987 stock-market break, results in more expensive caps, shown in Charts 2, 3, and 4.

Note: Gaps in Chart 1 result from missing observations.

Source: Chicago Mercantile Exchange.

expiration falls nearest to the futures expiration date.

The first example prices a two-year 10 percent cap, which consists of the sum of seven call options. At 10 percent, this cap is clearly out of the money. The computed call option price is expressed in basis points. The calls become progressively more expensive as the time to expiration increases, reflecting the rising time value of the calls. The shorter-maturity calls have little value because they are out of the money and, given the volatility, only a slight chance exists that they might finish in the money. Although the more distant calls are also out of the money, there is more time (and more uncertainty) about what LIBOR will do. Thus, their value is greater because of the higher probability that they might expire in the money. The sum of these calls is the cap rate, which is 147 basis points (rounded from 147.1).[15] For a three-month contract with a nominal face value of $1 million, a one-basis-point move is worth $25 ($1 million x .01% x 90/360). Translated into dollars, 147.1 basis points is $3,677.60 (147.1 x $25), which represents the dollar cost of placing a cap for two years on a $1 million loan. This example was computed ignoring the risk of default on the cap. It also assumes that payments at reset dates, if owed, are made at the time of the reset date.

Next, a slightly out-of-the-money 7.5 percent floor is shown. The total cost is 96 basis points, or $2,396.61. As mentioned above, the cost of interest-rate protection can be reduced by creating a collar, which is sometimes referred to as a ceiling-floor agreement. In this example,

Table 1.
Examples of Two-Year Cap, Floor, and Collar Prices on Three-Month LIBOR

Panel A: June 19, 1989; Volatility, 18 percent

	September 1988	December 1988	March 1989	June 1989	September 1989	December 1989	March 1990
Time to expiration (days)	91	182	273	364	455	546	637
Forward rate	9.02	8.84	8.64	8.71	8.77	8.87	8.86
Risk-free rate	8.46	8.47	8.54	8.56	8.59	8.59	8.56
Call prices (10.0 percent strike)	5.3	10.3	12.9	19.9	26.5	34.1	38.1
Put prices (7.5 percent strike)	.6	4.7	11.8	15.4	18.6	20.6	24.2

10 percent cap	7.5 percent floor	Zero-cost collar
Cost in basis points: 147	Cost in basis points: 96	10 percent cap implies
Cost in dollars: $3,677.60	Cost in dollars: $2,396.61	7.85 percent floor

Panel B: June 19, 1989; Volatility, 18 percent

	September 1988	December 1988	March 1989	June 1989	September 1989	December 1989	March 1990
Call prices (11 percent strike)	.4	2.2	3.8	7.6	11.8	16.9	20.3
Put prices (7 percent strike)	.1	1.3	4.7	7.2	9.5	11.2	13.9

11 percent cap	7 percent floor	Zero-cost collar
Cost in basis points: 63	Cost in basis points: 48	11 percent cap implies
Cost in dollars: $1,575.84	Cost in dollars: $1,198.08	7.19 percent floor

Panel C: December 14, 1987; Volatility, 25 percent

	March 1988	June 1988	September 1988	December 1988	March 1989	June 1989	September 1989
Time to expiration (days)	91	182	280	371	455	553	644
Forward rate	8.09	8.34	8.62	8.88	9.11	9.31	9.48
Risk-free rate	6.09	6.79	7.11	7.51	7.66	7.79	7.92
Call prices (10 percent strike)	2.1	12.5	28.9	45.9	62.0	78.0	91.6
Put prices (7.5 percent strike)	16.2	23.0	26.8	29.0	30.5	32.9	34.8

10 percent cap	7.5 percent floor	Zero-cost collar
Cost in basis points: 321	Cost in basis points: 193	10 percent cap implies
Cost in dollars: $8,025.53	Cost in dollars: $4,829.68	8.05 percent floor

Note: Dollar amount is for $1,000,000 in notional principal.

selling a 7.5 percent floor would substantially reduce the cost of a 10 percent cap. The combination would cost about 51 basis points, or $1,281. However, by judiciously selecting the floor level—in this case, 7.85 percent—the price of the cap can be driven to zero.[16] Marketing people delight in explaining that downside interest-rate protection (the cap) can be obtained at no cost: just sell a floor.[17] Of course, though, this strategy carries a cost. The holder of an interest-rate collar has traded away potential savings on interest-rate declines below the floor. This caveat notwithstanding, a collar for which the floor exactly matches the cap will be referred to as a *zero-cost* collar.

Panel B illustrates how the cost of caps and floors falls by selecting more out-of-the-money levels. Increasing the cap by one percentage point to 11 percent reduces the cap rate substantially to 63 basis points, or $1,575.84. Decreasing the floor by half a percentage point to 7 percent more than halves the cost to 48 basis points, or $1,198.08. A zero-cost collar with an 11 percent cap effectively lowers the floor to 7.19 percent.

The final example, reflected in panel C of Table 1, shows prices for caps, collars, and floors during the relatively high volatility period after the October 1987 stock market break. As depicted in Chart 1, Eurodollar futures' volatility surged during and after the October 21 crash; the degree of fluctuation had abated greatly by late January, although it had not returned completely to precrash levels. The implied volatility was 25 percent on December 14, 1987, as compared to 18 percent on June 19, 1989, in the earlier examples. The 10 percent cap priced in panel C is substantially more costly than the one in panel A. The cost is 321 basis points, or $8,025.53. Another important factor contributing to the higher cost is the rising structure of LIBOR forward rates. Although the futures nearest to expiration indicate a forward rate of 8.09 percent as compared to 9.02 percent in the June 19, 1989, example, the distant futures for December 14, 1987, have forward rates that are well above those for June 19. The upward sloping term structure of interest rates for December 14 reinforces the effect of higher volatility on raising cap and floor rates. The floor is more expensive as well at 193 basis points, or $4,829.68. Interestingly, the zero-cost collar with a 10 percent cap is only slightly more constraining with a floor of 8.05 percent as compared to 7.85 percent in the previous example, which exhibited low volatility and flat term structure.[18]

Caps, Collars, and Floors in Practice

At first sight, creating caps, collars, and floors would appear to be a simple matter because options are traded on the Eurodollar futures contract. Selecting the appropriate strike levels and expiration dates would appear to be all one needs to manufacture a cap, collar, or floor. However, as mentioned above, Eurodollar contracts extend into the future for at most four years (which nevertheless is an unusually large number of months for a futures contract). Eurodollar futures options traded at the Chicago Mercantile Exchange currently have expiration dates ranging out only two years, in a quarterly cycle that matches that of the Eurodollar futures contracts.[19]

Another limitation of Eurodollar futures options is that only contracts expiring within the three months or so from the current date are liquid, that is, they are the only ones that are actively traded so that their prices at any time reliably reflect equilibrium values. The options also are limited to strike levels in increments of 25 basis points, whereas the futures have increments of one basis point. Unlike Eurodollar futures and options, caps, collars, and floors have been created with maturities extending as much as 10 years. Furthermore, actual caps, collars, and floors can be created on any day, not just on futures and options expiration dates. The actual use of futures and options to fashion caps, collars, and floors is neither a straightforward nor a riskless matter.

The solution to this problem is the use of existing futures and options contracts to create the desired positions synthetically. Synthesizing an options position using options or futures contracts—or a combination of the two—requires not only taking appropriate positions in the existing liquid contracts but also altering that position over time so that the value of the actual position tracks or "replicates" the desired position. This process is known as *dynamic hedging*. Theoretically, the replicating portfolio of actual futures and options contracts can exactly match the value of, say, a cap sold to a counterparty.[20] In reality, managing a replicating portfolio is a risky and costly activity.[21] Tracking errors cumulate since costly trading cannot be conducted continuously as is theoretically required and because mismatches can occur with the expiration dates and possibly also with the interest rates involved. Using Eurodollar futures to hedge a cap based on the commercial paper rate exemplifies the latter.[22]

The Over-the-Counter Market

In view of the complexities and risks of dynamic-hedging strategies, most cap, collar, and floor users prefer over-the-counter instruments. Commercial and investment banks create these instruments themselves, possibly by manufacturing them through dynamic hedging. Nonfinancial users tend to rely on the expertise of these financial institutions and are willing to pay for the convenience of interest-rate risk management products issued through an intermediary. The intermediaries may also be more willing to bear the risks associated with hedging because of the scale of their operations. In fact, Keith C. Brown and Donald J. Smith (1988) describe the increasing involvement of banks in offering interest-rate risk management instruments as the reintermediation of commercial banking. Since the 1970s, commercial banks have played less of a role in channeling funds from lenders to borrowers. With the growth of interest-rate risk management, though, their intermediary role is being restored, albeit in a different form.

Commercial banks, particularly the largest money-center banks, are better able to absorb and control the hedging risks associated with managing a caps, collars, and floors portfolio, and these institutions are better able to evaluate the credit risks inherent in instruments bought from other parties. Credit risk arises because any counterparty selling a cap, for example, is obligated to make payments if the cap moves in the money on a reset date. That counterparty could go bankrupt at some point during the course of the cap agreement and would default on its obligation. (This issue is examined in detail in the last section of this article.) By taking positions in caps, collars, and floors, commercial banks—and to a lesser extent, investment banks—act as dealers by buying and selling to any counterparties. Within their portfolio or "book" of caps and floors, individual instruments partially net out, leaving a residual exposed position that the banks then hedge in the options and futures markets. Much trading of caps, collars, and floors consists of purchases and sales of these instruments to adjust positions and risk exposures, so much of the caps market's volume is generated by interdealer transactions. In addition to the dozen or so commercial and investment banks in New York and London that dominate the caps market, there are about half a dozen caps brokers, who do not take positions themselves but instead match buyer and seller.[23]

Caps, collars, and floors are usually sold in multiples of $5 million, but because of the customized nature of the over-the-counter market other amounts can be arranged. Most caps have terms that range from one to five years and have reset dates or frequencies that are usually monthly, quarterly, or semiannual. Caps based on three-month LIBOR are the most common and the most liquid or tradable. From the purchaser's point of view, buying a cap that matches the characteristics of the liability being hedged might seem best. Even strike levels and notional principal amounts can be chosen to vary over the term of an agreement in a predetermined way, but good fit comes at a price. Transactions costs are higher for such tailored products, as reflected by the larger difference between bid and offer rates on uncommon caps. This wider spread also increases the cost of removing caps by selling them before their term expires. Many users opt for a liquid cap and are willing to absorb the basis risk—the risk from a mismatch of interest basis or other characteristics—in order to avoid the higher cost of a less liquid instrument.

Caps and floors are usually available at strike levels within several percentage points of the current interest-rate basis and are most commonly written out of the money. Settlement dates typically occur after reset dates, upon maturity of the underlying instrument. For example, interest on a three-month Eurodollar deposit is credited upon maturity of the deposit. A cap on three-month LIBOR would have a three-month lag between a reset date and actual settlement. Most payments for caps are made up front, although they can also be amortized. When a cap and a floating-rate loan come from the same institution, the two are usually treated as a single instrument; thus, when the floating rate exceeds the strike level, payment is limited to the strike level and the cap does not pay off directly.[24]

Long-Term Caps. During the mid-1980s, early in the development of the caps market, longer-term caps were created directly from floating-rate securities rather than synthetically. Two kinds of floating-rate instruments were used: floating-rate CDs and floating-rate notes.[25] Floating-rate notes are debt obligations usually indexed to LIBOR, and floating-rate CDs are medium-term deposit instruments that are also typically indexed to LIBOR. The innovation that sparked much activity in the caps market was the issuance of capped floating-rate notes and CDs that in turn had their caps stripped off and sold as separate instruments sometimes known as "free-standing" caps.

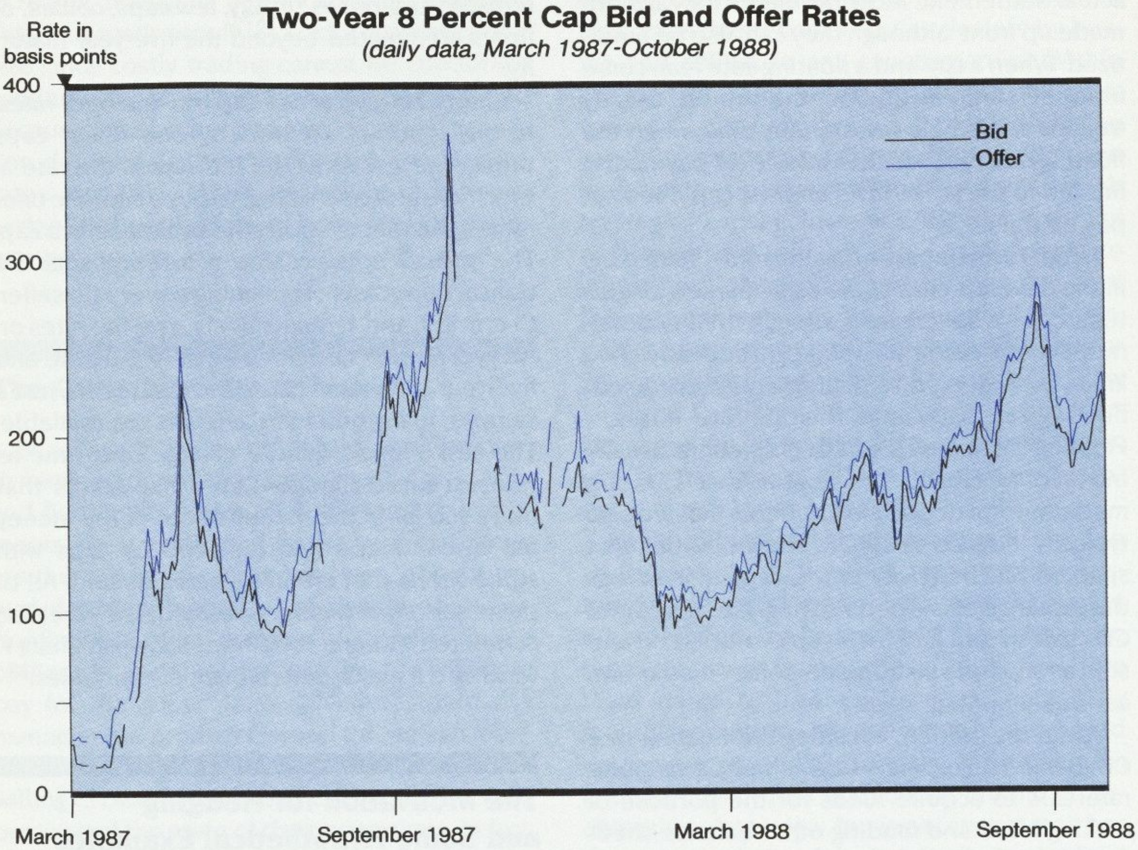
As an illustration, consider the floating-rate CD. Banks use ordinary CDs as well as variable-rate CDs to acquire funds for the purpose of making loans and funding other balance-sheet assets. The capped floating-rate CD was promoted as a method of raising funds below LIBOR, the rate on an uncapped CD with a variable rate of interest. The reason is that, after issuing a capped floating-rate CD to a depositor, a bank could then sell the corresponding cap into the caps market and collect premium income. Because CDs of this type typically fund floating-rate loans, the bank would be fully hedged after selling the cap. Funding costs would be lowered if the premium for the cap on the floating-rate CD were less than the premium that the bank collected upon selling the cap into the market.[26] This method of creating or "sourcing" caps, floors, and collars—through capped floating-rate CDs and floating-rate notes—became extremely popular but was short-lived. Reportedly, the longer-term caps were gradually perceived to be undervalued, such that cap writers were not being compensated for the risks of having to make payments to cap holders if interest rates rose above strike levels.[27] Also contributing to the demise of this method of sourcing was a flattening of the yield curve that made floating-rate borrowing less attractive and reduced cap prices. Today, few caps, collars, or floors are created beyond the five-year maturity.

Charts 2-4 give actual cap bid and offer rates, in basis points, quoted by one major caps broker in New York. The bid rate is the rate at which the broker is willing to buy a cap; the offer rate is the rate at which the broker sells a cap. The spread between the two represents the transactions costs of matching buyer with seller. Charts 2, 3, and 4, respectively, give the rates on two-year 8 percent, three-year 10 percent, and five-year 10 percent caps. These rates are just a sample; many other strike levels are available. The strike levels quoted change over time as interest rates change. Cap strike levels that move too far in the money or out of the money are discontinued and replaced by caps with strike levels that are in greater demand. All of these series are highly correlated. They are also correlated with the volatilities shown in Chart 1, which are a major determinant of cap values.[28]

The Motivation for Hedging and Some Hypothetical Examples

With some background on the caps, collars, and floors market, the use of interest-rate risk management instruments can now be put into perspective by briefly considering the nature of hedging. Caps, collars, and floors are often talked about in terms of an insurance analogy. They are instruments that can be used to hedge assets or liabilities and thus protect against loss resulting from interest-rate risk. In practice, though, distinguishing between hedging and speculating in interest-rate risk management is sometimes difficult, especially with option-based instruments. Discretion is required in selecting the timing of the hedge, the strike level, and the maturity of the instrument, all of which are usually predicated on some opinion of what interest rates and other variables are expected to do. Selling a cap or floor, for example, is a way to generate income on a fixed-income portfolio by collecting the premiums. The decision to sell often reflects a difference of opinion regarding the volatility implied by the

Chart 2.
Two-Year 8 Percent Cap Bid and Offer Rates
(daily data, March 1987-October 1988)



The spread between the bid and offer rates represents the transactions costs of matching buyer and seller.

Note: Gaps in Charts 2, 3, and 4 reflect days for which rates were not available.

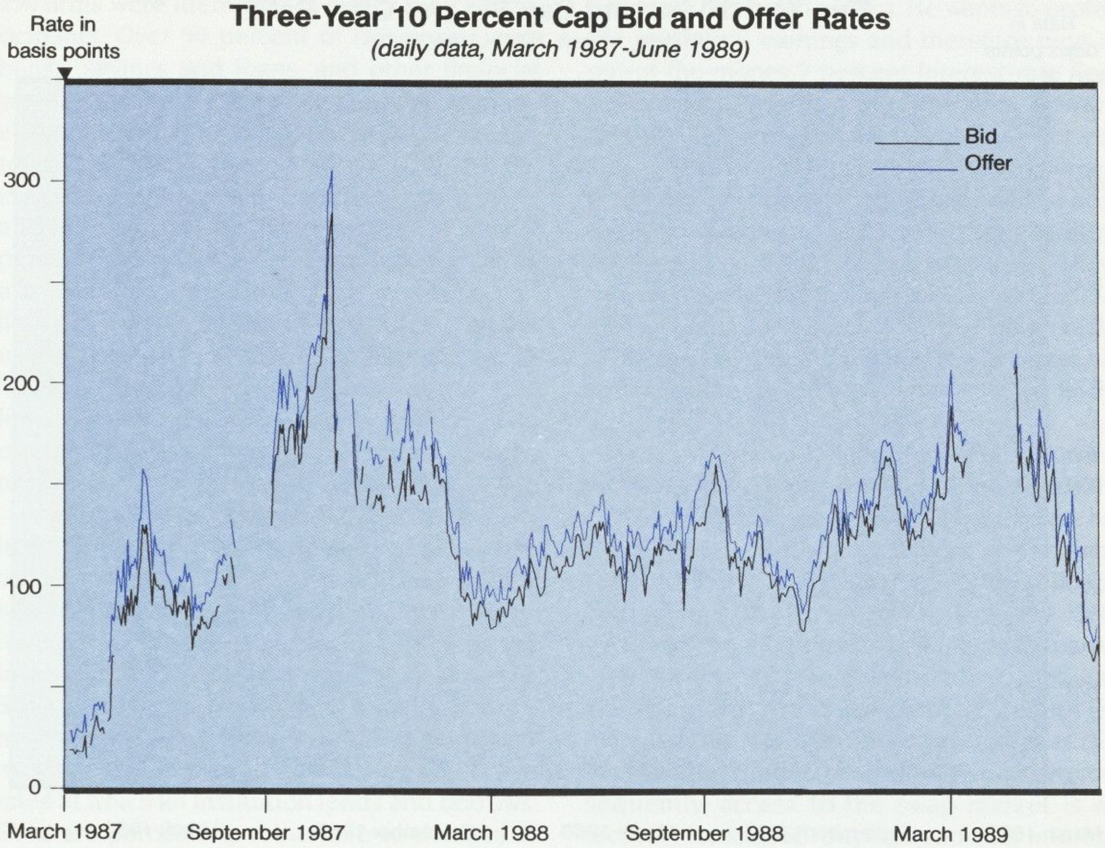
Source: Noonan, Astley, and Pearce, Inc.

cap or floor. If a money manager thinks a cap is overvalued because the market's expectation of volatility is higher than his or her own, then selling an out-of-the-money cap might be a good move. If the money manager's judgment about volatility is correct, even small upward moves in the interest rate may not wipe out all of the premium income. At the same time, the sale provides a limited hedge against small downward moves in rates, again because of the premium receipt.

Even determining the effect of hedging can be problematic, since a firm's purchase of a cap, for example, to hedge the interest-rate risk of a particular liability could increase the variability of the firm's net worth. The financial claim being hedged may itself help offset the variability of another financial claim on the balance sheet. The net result of a specific hedge could be to increase the interest-rate risk exposure of the firm.

A more fundamental issue is why firms hedge in the first place. A basic insight derived from the economics of uncertainty is that risk aversion leads individuals to prefer stable income and consumption streams to highly variable ones. Given an assumption of risk aversion on the part of decision makers, one can show that their welfare or utility (that is, their economic well-being) is greater over time if they enjoy smooth income or consumption opportunities rather than erratic ones.[29] Hedging is a way of improving economic well-being by trading off income or consumption in good times for greater income or consumption in bad times. Thus, a hedging strategy serves a well-defined purpose

Chart 3.
Three-Year 10 Percent Cap Bid and Offer Rates
(daily data, March 1987-June 1989)



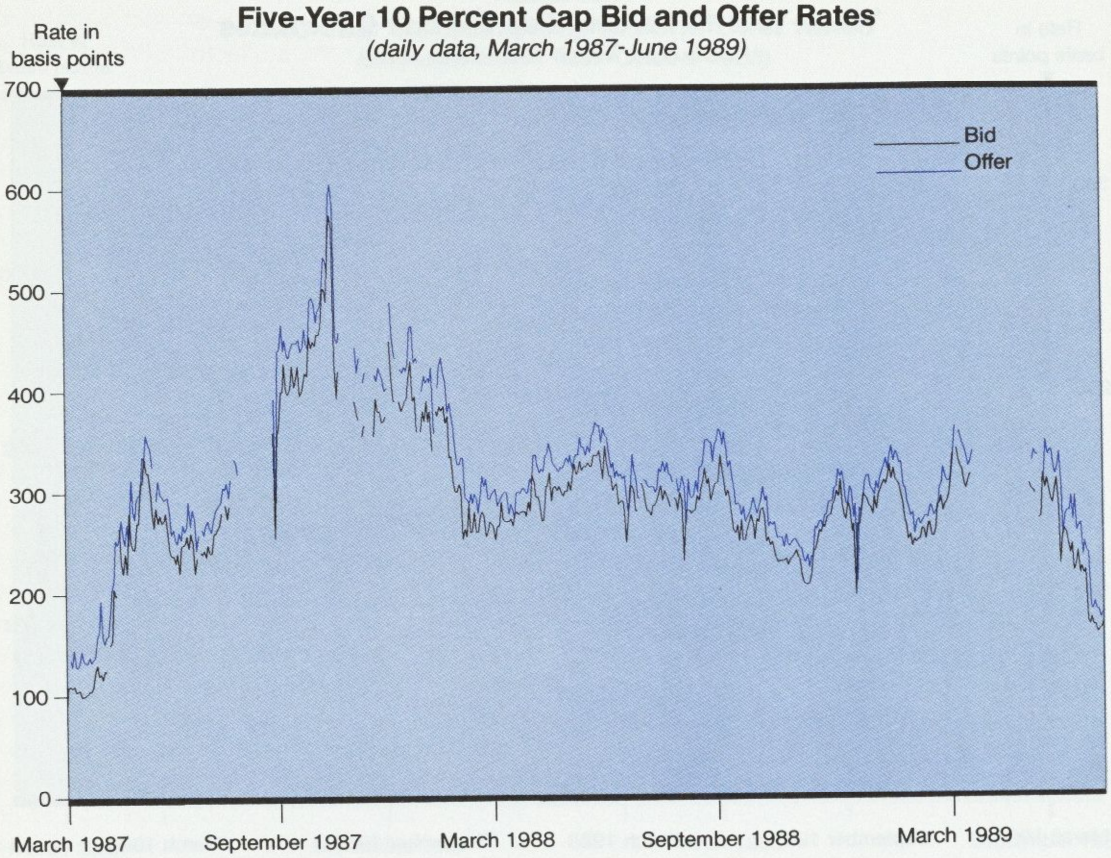
The rates depicted in Chart 3 are highly correlated with those in Charts 2 and 4, as well as with the volatilities in Chart 1.

Source: Noonan, Astley, and Pearce, Inc.

for risk-averse economic agents, such as farmers or a firm's owner-manager. The issue is less clear-cut for widely held corporations, which actually are the typical users of interest-rate risk-management tools. A corporation owned by a large number of stockholders need not operate like a risk-averse decision maker because each stockholder can insulate his or her wealth and consumption opportunities from risk, specific to the corporation's activities, by holding a diversified portfolio of assets.

Clifford W. Smith and René M. Stulz (1985) surveyed managers of widely held, value-maximizing corporations to determine the motivations behind hedging behavior. According to the researchers, managers engage in hedging of a firm's value for three basic reasons. The first explanation is tax-related; Smith and Stulz argue that, on average, a less variable pretax firm value implies a higher after-tax firm value than does a more variable pretax value. The reasoning turns on their assumption that the level of corporate tax liabilities grows at an increasing rate with rising pretax firm value because of the progressive structure of the tax code. Hedging helps reduce the variability of pretax firm value and therefore raises after-tax value. Second, Smith and Stulz maintain that hedging lowers the probability that the firm will go bankrupt and thus incur bankruptcy costs. Hedging firm value would benefit stockholders by reducing the expected future costs of bankruptcy that lower current firm value. A related point is that a firm's debt may often contain covenants that force the company to alter investment policies that the shareholders would like to see under-

Chart 4.
Five-Year 10 Percent Cap Bid and Offer Rates
(daily data, March 1987-June 1989)



The longer expiration date on a five-year cap results in prices that are relatively higher than those on shorter-term caps.

Source: Noonan, Astley, and Pearce, Inc.

taken. Hedging reduces the likelihood of financial distress and the limitations on managers' discretion that bond covenants may impose. A third reason for hedging is that when managerial compensation is tied to the firm's value, managers may become more risk-averse in order to maintain that value.

Participants in the Caps, Collars, and Floors Market. While the precise social value of interest-rate risk management products is not fully understood in the case of widely held corporations, such products are clearly becoming increasingly popular among corporate treasurers and other financial managers. End users of caps, collars, and floors typically include firms seeking to limit exposure to adverse movements in short-term interest rates, such as a firm that sells commercial paper to fund its purchases of inventory. Specific market participants are depository institutions, particularly savings and loan associations (S&Ls); corporations going through leveraged buyouts (LBOs) or taking on debt to fend off hostile takeovers; and real estate developers, who are often highly leveraged with floating-rate debt. Unfortunately, the only information about these applications is anecdotal. Also, compared to the potential market, the actual market is probably very small. Many potential users are unaware of or cautious about interest-rate risk management instruments.

Any user of interest-rate swaps is potentially also a user of caps, collars, and floors. Larry D. Wall and John J. Pringle (1988) conducted a systematic search of annual reports for 4,000 firms that used interest-rate swaps in 1986. The stocks of these firms were traded on the New York

Stock Exchange, the American Stock Exchange, or the over-the-counter market. Of this sample, 250 firms were identified as swaps market participants. Over 50 percent of this group were banks, savings and loans, and other financial services firms; commercial banks alone accounted for half of these. In addition, Wall and Pringle report that "the overwhelming majority of thrifts (59 percent), manufacturing firms (69 percent), and nonfinancial, nonmanufacturing firms (77 percent) are exclusively fixed-rate payers."[30] As a conjecture, the profile of caps, collars, and floor users may be quite similar to that for swaps users. The fact that credit risks for caps and floors are one-sided, however, suggests that firms with weaker credit ratings probably use caps and floors because they cannot gain access to the swaps market on favorable terms.

Anecdotal accounts from various sources illustrate how different end users employ caps, collars, or floors in their management of interest-rate risk. Many savings and loans, for instance, have been active users of these option-based instruments. The interest-rate risk confronting S&Ls, and depository institutions generally, may be considered in terms of their net interest margins, that is, the difference between the rates at which an institution lends and borrows. S&Ls are particularly vulnerable to changes in interest rates because maturities (or alternatively, the durations) of these institutions' assets, predominantly long-term mortgages, greatly exceed the maturities of their liabilities, most often short-term time and savings deposits. Thus, a rise in rates raises the interest expense on an S&L's short-term liabilities with possibly little increase in interest earnings on its mortgages. The net interest margin narrows and could very well become negative. One solution is to convert the floating-rate interest expense on the liabilities into fixed-rate payments via an interest-rate swap. The net interest margin would then become much more stable. However, a weak credit standing could make such a swap too expensive or unobtainable. A cap on the floating-rate liabilities could be an effective alternative. An S&L's credit rating would be irrelevant to a cap writer, who bears no credit exposure.[31]

As another example, consider a commercial bank's portfolio manager who is responsible for overseeing a portfolio of floating-rate notes. Suppose this manager believes that a large drop in short-term interest rates, currently at about 8 percent, is about to occur. He wants to protect the portfolio's earnings and therefore buys an out-of-the-money 7 percent interest-rate floor. Concerned about the cost of this protection and reasonably convinced that rates will not rise substantially, he also decides to sell a 9 percent interest-rate cap to create a collar on the portfolio. This example highlights the discretion involved in selecting a hedge. A floor could have been in place all along, but maintaining a floor reduces a portfolio's return by the amount of the premium expense. Only when the manager has strong concerns about a drop in rates is the floor purchased.

As a final example, the corporate treasurer of a consumer products firm is worried about the prospects of a rise in interest rates because her company has recently undergone a leveraged buyout. The financing strategy for the LBO included heavy reliance on floating-rate debt secured from a syndicate of commercial banks. The firm's debt-to-equity ratio has soared, and even a modest rise in rates could bankrupt the company. After the LBO the firm's credit standing was downgraded by the rating services; consequently, access to the swap market is effectively foreclosed. Buying a two-year interest-rate cap to cover the firm's floating-rate exposure seems to be a prudent action.[32] The treasurer expects earnings will be more robust after a two-year interval. Also, the protection gained for a relatively short-term horizon makes sense because during this period the firm would be downsizing and reorganizing its operations.

Credit Risk

The earlier discussion of the over-the-counter market for caps, collars, and floors alluded to the risk of default inherent in these instruments. That risk is present because the seller of a cap or floor is agreeing to fulfill a contract in the event the cap or floor moves in the money on a payment date. Since the seller is a firm, whether a commercial bank, investment bank, or nonfinancial institution, its assets are limited, and thus the company is exposed to the possibility of bankruptcy. The probability of default is rather small for the typical caps, collar, or floor writer who also typically issues investment-grade bonds into the market. Moody's Investors Service, one of the major bond rating firms, recently released a study indicating that from 1970 to 1988 the average annual rate of default by issuers of investment-grade bonds was 0.06 percent, as compared to an average annual default rate of 3.3 percent for junk bond issuers.[33] Because the consequences of default can be financially damaging, default risk receives careful analysis, particularly by counterparties entering into caps and swaps agreements. This section of the article takes a detailed look at how default risk is evaluated and how it affects the pricing of caps, collars, and floors.

The first aspect of the problem is to consider the precise nature of the default risk or, alternatively, the credit exposure. If a cap is in the money on a floating-rate reset date, the owner of the cap expects to receive a payment from the cap writer, as reviewed above. If the writer is insolvent and thus fails to make the payment, the owner is again in an unhedged position and must make the full floating-rate payment, but this is not the only consequence of default. Provided the default does not occur on the final reset date, the cap was also hedging future reset dates, which upon default are also fully exposed. Thus, credit exposure depends on the time that default occurs in the life of a cap agreement. (Note that a parallel argument can be made for floors and collars.) The cost of default to the cap buyer is the cost of replacing the original cap with a new cap from another seller. If interest rates at the default date were identical to the initial interest rates and the volatility had not changed since the original cap was purchased, the replacement cost of the cap would be zero, ignoring transactions costs and differences in credit risks. That is, the cost of a new cap for the remaining reset dates would exactly equal the current market value of the existing cap (if default had not occurred).

The next and rather complex aspect of the credit risk question to consider concerns the method of assessing credit risk when a cap is sold. Bankruptcy of a cap writer has no impact on cap buyers as long as the cap stays out of the money and the cap buyer has no intention of selling the cap before its term ends. Default occurs only when a cap is in the money and the cap writer is bankrupt. The likelihood of bankruptcy may also be related to the level of interest rates and thus dependent on the future path of these rate movements. In addition, as just discussed, a cap's replacement cost is a function of where in the life of the cap agreement default occurs. All of these factors should be weighed in evaluating what the potential cost of default could be and how that should affect the price of a cap.

Marcelle Arak, Laurie S. Goodman, and Arthur Rones (1986) propose a method of computing credit exposure for caps, collars, and floors. Their approach amounts to considering different worst-case scenarios that are defined by the degree to which a cap can move in the money. For a cap the computed exposure depends on the size of the upward movement in the interest rate that could occur during each reset interval. A cap's replacement value will tend at first to increase early in the life of the instrument and then to decrease toward the end of the contract. The credit exposure is taken to be the maximum replacement value computed at the reset dates. For example, if the interest-rate volatility based on three-month LIBOR is 10 percent (as measured by the annual standard deviation), over a three-month period the volatility is $0.10 \times \sqrt{(1/4)} = 5$ percent.[34] Assuming an initial 7 percent LIBOR, three months later the upward move would be to 7.35 $[7.0 + (0.05 \times 7.0)]$. Given this rate and a further assumption that rates at all other maturities shifted in parallel, the cap replacement value is calculated. Another 5 percent upward move is then computed, giving a new LIBOR of 7.72 $[7.35 + (7.35 \times 0.05)]$ and again the replacement

"The cost of default to the cap buyer is the cost of replacing the original cap with a new cap from another seller."

value is computed, and so forth. The credit exposure is the maximum value of the replacement cost during the cap agreement.

A more conservative evaluation of credit exposure might assume that rates rose by two standard deviations per year instead of one as in the previous example. At two-standard-deviation moves, the actual exposure would, on average, exceed the maximum computed amount only 2.5 percent of the time (as compared to exceeding it 16.5 percent of the time using a one-standard-deviation measure).[35] Arak, Goodman, and Rones give an example of the credit exposure on various collar agreements with a floor equal to 6 percent and a cap equal to 9 percent. For three-month reset intervals, the exposure is 0.44 percent of the notional principal (two-year collar), 0.82 percent (five-year collar), and 2.68 percent (10-year collar).[36] By these researchers' calculations, the credit exposure on collars is rather small, especially compared to similar calculations for other instruments they consider, such as interest-rate swaps and forward contracts. These calculations are intended for commercial banks, which set credit limits for particular customers in order to manage the size of potential losses in the event of default. However, the method put forth by Arak, Goodman, and Rones is not useful for pricing caps—that is, for adjusting the price or rate for the anticipated cost of default. Computations based on worst-case scenarios implicitly overstate the actual incidence of default because of the arbitrary assumption about sequential interest-rate moves only in one direction. A more desirable approach would compute the "expected value" of default—the difference between caps not subject to default and those that are.

"Computations based on worst-case scenarios implicitly overstate the actual incidence of default because of the arbitrary assumption about sequential interest-rate moves only in one direction."

Caps as Default-Risky Options. Almost all of the option pricing models used to value caps ignore default risk. An exception is the model proposed by Herb Johnson and Stulz (1987), in which they derive formulas for default-risky or "vulnerable" puts and calls. Unfortunately, their formulas cannot be straightforwardly applied to caps, collars, or floors because of the time dimension involved in these options-based instruments. As has been emphasized, caps are a sequence of options—default-risky options. Fulfilling a given option contained in a cap depends on the absence of bankruptcy at earlier reset dates. If bankruptcy occurred earlier, the current option would not be honored by the cap writer. The sequential time dimension involved in valuing caps makes the mathematics formidably complex.[37]

This author has tackled the complexity of cap valuation by using computer-intensive methods to handle the intricate contingencies implied in cap, collar, floor, and swap agreements (Peter A. Abken [forthcoming]). His computer model avoids the contradictory assumption inherent in the Black model used for short-term debt options—that short-term interest rates are constant—but at the cost of trading off a simple analytical formula for a complicated computer algorithm. Nevertheless, the intuition behind the new model is simple and easily explained.

The value of a European option can be thought of as the average or expected value of its payoffs at expiration, discounted back to the present. Options are difficult to value because the payoff upon expiration is a "kinked," or discontinuous, function of the underlying asset price. A call option is worth zero if the underlying asset price at expiration is less than the strike price, and positive in value if the underlying asset price exceeds the strike price, increasing dollar for dollar with the amount above the strike. The Black-Scholes and Black formulas compute the value of a call as the expected value of the future payoffs.[38] Some payoffs are more likely to occur than others, and the formulas account for the probabilities associated with the payoffs.

Monte Carlo Simulation. One method for valuing options relies on extensive computations to determine the expected payoffs. Known as Monte Carlo simulation, this process was first applied to option pricing problems by Phelim P. Boyle (1977). The standard application involves stock option pricing. A stock price, on which an option is valued, is assumed to rise and fall randomly over time, although its value at any point can be described in terms of its statistical distribution, which is known or assumed. In standard problems the distribution for stock price changes is assumed to be fully characterized by its mean and variance. Using this information, artificial future stock-price paths, also known as *realizations*, can be created numerically by computer. By randomly generating a large enough number of price paths (tens of thousands, at a minimum) and evaluating the payoff on an option with a given strike price at a particular point in time—the option's expiration date—an average over these randomly generated payoffs can be made. The option price is given by appropriately discounting the expected future payoff into current dollars. Of course, the Black-Scholes formula accomplishes the same thing mathematically and is conceptually equivalent. To the penny, both methods will give the same price using identical assumptions regarding the statistical characteristics of stock price movements. The Monte Carlo method, though cumbersome, pays off in cases where the asset price moves in unusual ways, such as in random jumps—for example, due to a stock market crash. The Black-Scholes model rules out such movements by assumption. Cap valuation is another area where Monte Carlo methods offer a simplification over approaches that may not otherwise be mathematically tractable.

Three factors taken together contribute to the complexity of default-risky cap valuation. The first is that debt prices on instruments like Treasury bills or Eurodollar deposits vary with interest rates. Second, each constituent option in a cap is subject to default and must be valued as a default-risky option. Third, the payoff on a given option depends on the nonoccurrence of default on options from earlier periods.

The payoff of a vulnerable call option is the lesser of the firm's value or the default-free option payoff. The value of the firm is the market value of its equity (before including the value of its cap). If the value of the firm that sold the option is greater than the payoff, no default occurs. If the payoff exceeds the firm's value, the company defaults and the option holder receives the value of the firm—or some share of it, as determined by the bankruptcy courts—when the company is liquidated. In view of the fact that a vulnerable call may pay off less, but never more, than a default-free call, the value of a vulnerable call must be less than the value of an otherwise comparable default-free call.

An additional consideration for cap valuation, as discussed above, is that default on a cap leaves the cap buyer unhedged. The exposure is the replacement value of the cap. Thus, default involves at a minimum replacement of the missed option payoff, and possibly the entire remaining value of the cap, if the firm wants to maintain the hedge. Thus, besides valuing default-risky call options, cap valuation must also evaluate such replacement costs.

The Elements of the Caps Model. To convey the basic ideas behind construction of the caps model, this section of the article sketches out the model, the technical details of which can be found in Abken (forthcoming). Three so-called state variables are computer-generated to implement the simulation. The options making up a cap are valued based on the underlying interest rate, as discussed earlier. The entire path of the term structure of interest rates is generated using the model developed by Stephen M. Schaefer and Eduardo S. Schwartz (1984). Two state variables are the difference or spread between the instantaneous rate and a consol rate (that is, the rate on a bond having infinite maturity), and the consol rate itself. All other intermediate-maturity discount bonds are derived by formula from these two inputs, which describe absolute and relative movements in interest rates at all maturities. The third state variable represents the value of the firm, which also fluctuates randomly over time, reflecting unpredictable changes in interest rates, earn-

"An additional consideration for cap valuation . . . is that default on a cap leaves the cap buyer unhedged. The exposure is the replacement value of the cap."

ings, and other variables that determine firm value.

The example to be considered is parallel to the one discussed earlier in Table 1, but the focus is now on credit risk. The cap model will value two-year caps on a three-month interest rate. The cap consists of seven reset dates, at each of which the firm's value is compared to the call option payoff. Default-free and default-risky caps are valued. The difference in the price or rate for these otherwise identical caps is the credit spread for default risk. The example developed below illustrates how default risk is particularly sensitive to the correlation over time between firm value and interest-rate movements.

The parameter values for the Schaefer-Schwartz model were estimated from actual interest-rate data on one-month Treasury bill and 30-year bond yields, which served as proxies for the instantaneous interest rate and consol interest rate, respectively. The rates were sampled weekly on Fridays from January 1983 to August 1989. The reader is referred to Abken (forthcoming) for details concerning parameter estimation and other technical details concerning the model.

"[D]efault risk is particularly sensitive to the correlation over time between firm value and interest-rate movements."

A simplification used in the simulations presented in this article is that whenever a default occurs—that is, when the firm value is less than the option payoff—the replacement value of the cap is not computed. Instead, the option payoff for that reset date is set equal to the negative of the payoff. In other words, the cap owner has to cover the full floating-rate interest payment for that date. Payoffs at future reset dates are assumed to be zero. Valuing a new cap at current rates would increase the cost of default compared to the procedure used here; such valuation, however, would also require separate simulations at each occurrence of default.

More Examples. Table 2 gives the results of the simulations. Three panels of this table differ only in the degree that firm value is correlated with interest-rate movements. In the Schaefer-Schwartz model, there are two elements to this correlation. Firm value can be correlated with consol rate movements or spread movements, or both. (The Schaefer-Schwartz model assumes that the spread and consol rate are uncorrelated, which is supported by empirical research.) Correlation coefficients range from −1, perfect negative correlation, to 1, perfect positive correlation. Intuitively, a cap writer whose firm value is negatively correlated with interest-rate movements poses a greater credit risk than one that is positively correlated. For a given strike level, when interest rates are high, caps are more likely to be in the money and require a payment from the writer. A negative correlation therefore means that high interest rates are associated with low firm value; hence, default is more probable than it would be for zero or positive correlations. Also, empirically short- and long-term interest rates are positively correlated. Thus, a negative correlation of firm value and long-term interest rate would also be associated with a negative correlation between the firm value and interest-rate spread (defined as the short rate less the long rate).

Panel A gives the base case of zero correlation of firm value with the interest-rate spread and with the long-term interest rate. The annual default rate for this case is set to 0.13 percent by adjusting the initial value of the firm to give this rate as the outcome of the simulations.[39] The same initial firm value is then used in panels B and C, thereby yielding new default rates due to different correlations with the term structure variables. The initial term structure has a spread of 2.7 percentage points, which was the average spread over the sample period. The short-term interest rate is initially 8 percent and the cap is written at 9 percent. As in Table 1, the option rates are given for each reset date. This table includes default-free and default-risky options; the sum over reset dates for each type is the cap rate. Because the default rate is so low, the discrepancies between default-free and default-risky option prices do not become significant

Table 2.
Default-Free and Default-Risky Cap Rates Estimated by Monte Carlo Simulation, 9.0 Percent Two-Year Cap

Initial term structure: Short-term rate, 8.0 percent; Long-term rate, 10.7 percent

Panel A: Correlation of firm value with interest-rate spread: 0
Correlation of firm value with long-term rate: 0

Reset date number:	1	2	3	4	5	6	7
Time to expiration (weeks):	13	26	39	52	65	78	91
Default-free option rate:	7.94	17.99	26.95	35.57	43.98	51.87	59.86
Default-risky option rate:	7.94	17.99	26.95	35.53	43.81	51.34	58.71

Default-free cap rate:	244.16	Default-risky cap rate:	242.28
Standard deviation:	(1.45)	Standard deviation:	(1.43)
95 percent confidence interval:	(241.32, 247.00)	95 percent confidence interval:	(239.48, 245.08)

Credit spread in basis points:	1.89
Standard deviation:	(0.14)
95 percent confidence interval:	(1.62, 2.16)
Annual default rate:	0.13 percent

Panel B: Correlation of firm value with interest-rate spread: −0.5
Correlation of firm value with long-term rate: −0.5

Reset date number:	1	2	3	4	5	6	7
Default-free option rate:	7.94	17.99	26.95	35.57	43.98	51.87	59.86
Default-risky option rate:	7.94	17.98	26.79	34.98	42.19	48.15	53.51

Default-free cap rate:	244.16	Default-risky cap rate:	231.54
Standard deviation:	(1.45)	Standard deviation:	(1.35)
95 percent confidence interval:	(241.32, 247.00)	95 percent confidence interval:	(228.89, 234.19)

Credit spread in basis points:	12.63
Standard deviation:	(0.35)
95 percent confidence interval:	(11.94, 13.32)
Annual default rate:	0.71 percent

Panel C: Correlation of firm value with interest-rate spread: 0.5
Correlation of firm value with long-term rate: 0.5

Reset date number:	1	2	3	4	5	6	7
Default-free option rate:	7.94	17.99	26.95	35.57	43.98	51.87	59.86
Default-risky option rate:	7.94	17.99	26.95	35.57	43.98	51.87	59.86

Default-free cap rate:	244.16	Default-risky cap rate:	244.16
Standard deviation:	(1.45)	Standard deviation:	(1.45)
95 percent confidence interval:	(241.32, 247.00)	95 percent confidence interval:	(241.32, 247.00)

Credit spread in basis points:	0.0
Standard deviation:	(0.0)
95 percent confidence interval:	(0.0, 0.0)
Annual default rate:	0.0 percent

Note: Sample size for each panel: 50,000 independent draws. Cap rates expressed in basis points.

until the later reset dates. The default-free cap rate is 244.16 basis points, whereas the default-risky cap rate is 242.28. The difference of 1.89 basis points is the credit spread.

These figures are estimates and have an error associated with them. One can arbitrarily reduce that error by increasing the number of realizations used to compute the options. Quadrupling the number of realizations reduces the standard deviation by half. The simulations for each panel were generated by taking 50,000 independent sets of realizations of the state variables.[40] The standard deviation and 95 percent confidence intervals for each cap rate and the spread are reported in Table 2.

The simulation used to generate panel B was the same as that for panel A in all respects except that the correlation between firm value and interest-rate spread is −0.5 instead of zero, and the correlation between firm value and long-term rate is −0.5 instead of zero. The results show a substantial increase in the incidence of default. The base rate in panel A for zero correlations is 0.13 percent, whereas the negative correlations in panel B raise the default rate to 0.71 percent. The credit spread rises from 1.89 basis points to 12.63 basis points. As discussed previously, the reason is that firm value is likely to be low when interest rates are high. The cap writer has a greater chance of being insolvent when a payment is required. As a final example, the correlations in panel C take the opposite signs from those in panel B. The credit spread and annual default rate drop to zero. The greater chance of high firm value coinciding with cap payments reduces the likelihood of default by the cap writer; in this case, the incidence of default drops to zero.

The substantial increase in the credit spread exhibited in panel B may exaggerate default risk for two reasons. First, the cap is assumed to be unhedged by the firm. In other words, the company is taking a speculative position. Actual cap writers usually take offsetting positions in other caps or hedge by other methods, at least to some degree. Second, the model assumes that failure to cover cap payments is the only factor causing bankruptcy. For actual cap writers, the contingent liability posed by a cap is probably small compared to other items on the balance sheet. On the other hand, the computed credit spread may still be a good approximation if the cap serves as a proxy for the firm's overall balance sheet exposure to movements in interest rates.

No data on actual credit spreads are published. In conversations with the author, cap market participants place the credit spreads that have occurred in the range of 5 to 10 basis points for two- to three-year caps. The estimated spreads using the cap model are roughly in that range. Further research into actual credit spreads and refinements of the cap model should sharpen the estimation results and make the model more useful.

Conclusion

Interest-rate caps, collars, and floors are among the newest interest-rate risk management instruments. This article has given an exposition of these closely related instruments, which are options-based and designed to limit exposure to fluctuations in short-term interest rates on floating-rate assets or debt. Their applications are not limited to hedging. Like options, they are also convenient for speculating on interest-rate movements. In practice, however, the distinction between these two applications is rarely clear-cut. Several examples served to illustrate how financial managers use caps, collars, and floors.

The article also discussed the credit risks associated with caps, collars, and floors, which for the most part are over-the-counter contracts offered by one firm to another. Default risk is inherent in this kind of arrangement and can be priced. A new cap valuation model produced credit spreads that are not much different from those observed in the cap market between stronger and weaker credit risks among cap writers. Interest-rate risk management has been growing in importance for financial managers. This article may improve their understanding of the credit risk of caps, collars, and floors and help determine the cost of interest-rate protection.

Notes

[1]Recent *Economic Review* articles include Abken (1987), Feinstein and Goetzmann (1988), Kawaller, Koch, and Koch (1988), and Feinstein (1989).

[2]See Wall and Pringle (1988) for an introduction to interest-rate swaps.

[3]For brevity, the market for caps, collars, and floors will be referred to as the *cap market.*

[4]See Kuprianov (1986): 16-20, for a discussion of Eurodollar deposits and Eurodollar futures.

[5]The information on the ISDA survey was reported in *Risk* 2 (April 1989): 11.

[6]A detailed discussion of option pricing is beyond the scope of this article. A basic overview can be found in Abken (1987). See Cox and Rubinstein (1985) or Jarrow and Rudd (1983) for more thorough introductions to option pricing.

[7]See Abken (1987): 6, for more detail.

[8]See Henderson (1986) for further discussion.

[9]See Black (1976).

[10]To the author's knowledge, no published studies have compared the accuracy of different option-pricing models for pricing caps and related instruments. One reason may be that there are no publicly available data on these rates, and another is that these instruments are relatively new. Little empirical research exists on the adequacy of different interest-rate option-pricing models. Boyle and Turnbull (1989) use the Courtadon option-pricing model in evaluating collar rates, but they do not compare their rates with those from other models nor with actual market rates.

[11]Because the CME and most LIFFE Eurodollar futures are "cash-settled," a $1 million deposit is rarely made, but instead only the difference between the current, or spot, LIBOR and the contracted LIBOR times the notional principal actually changes hands.

[12]Prior to June 1989 contract months extended three years.

[13]A Eurodollar futures price is actually an index value that equals 100 minus the "add-on" yield (three-month LIBOR). Thus, the futures price and add-on yield move inversely with each other. See Kuprianov (1986): 16, for more detail on Eurodollar futures and short-term interest-rate futures generally. Both the add-on yield and the futures price are usually quoted in the financial press.

[14]See Feinstein (1989) for details on the estimation, interpretation, and uses of implied volatilities. The Eurodollar futures options are actually American options, but the early exercise feature has negligible value for the slightly out-of-the-money options usually used in estimating the implied volatilities with a European futures option formula.

[15]Sums in Table 1 may not add up due to rounding error. Cap rates are usually rounded to whole basis points. The dollar amounts are the exact amounts computed in constructing Table 1.

[16]Another way to create a zero-cost collar is to set the floor first and then determine the appropriate cap. The method discussed in the text is more common.

[17]Collars have also been offered that give the buyer a payment for taking the collar, that is, the value of the floor sold exceeds the cost of the cap purchased. See "NatWest Uses Incentives to Push Rate Collars," *American Banker,* August 2, 1989.

[18]These examples are consistent with the recent findings of Boyle and Turnbull (1989) in their examination of collars. Using a different option-pricing model than the Black model, they found that a 100 percent increase in the volatility causes the floor level to change by less than one basis point. If their findings are also valid for the Black model, most of the difference observed in the examples in the text is attributable to the difference in yield curves.

[19]Before March 1989, contract expiration dates had a maximum maturity of one year. See Chicago Mercantile Exchange (February/March 1989): 7.

[20]The term *counterparty* is standard terminology for the other party in a swap, cap, floor, or collar agreement.

[21]Another complication in using futures in a replicating portfolio is that futures contracts are marked to market daily. This situation may create cash flow problems since futures positions that lose value may be subject to frequent margin calls. Even though the replicating portfolio is used to hedge a cap, which matches it in value, the cash flows from the cap come only when it is sold and on interest payment dates.

[22]See Abken (1987) for more on the synthetic creation of options. Mattu (1986) gives examples of replicating portfolios for caps and floors.

[23]Shirreff (1986) gives an interesting though somewhat dated overview of the caps market and the various players in it.

[24]LeGrand and Fertakis (1986): 134.

[25]Floating-rate CDs are also called variable-rate CDs.

[26]See *Intermarket* (October 1986): 14, for an account of the first such sale of a cap from a capped floating-rate note (FRN). By selling a cap off an issue of $100 million in 12-year capped FRNs, Banque Indosuez of Paris lowered its interest rate by one-eighth of a point below LIBOR. Uncapped, the notes would have sold at LIBOR. The capped FRNs were issued at LIBOR plus three-eighths. On an annual basis, Indosuez therefore collected the equivalent of 50 basis points on the sale of its cap.

[27]Shirreff (1986): 29.

[28]The volatilities shown in Chart 1 are probably not the same as those used to generate the cap rates. The volatilities were obtained from a different source than the cap rates, but they should be highly correlated with the actual volatilities used to price the caps.

[29]Newbery and Stiglitz (1981) give a comprehensive discussion of risk aversion and the rationale for hedging.

[30]Wall and Pringle (1988): 22.

[31]The example given was described in terms of a "flow concept" of interest-rate risk, that is, the impact of a change in interest rates on the net interest margin. Another way to view interest-rate risk is in terms of a "stock concept," the change in the net worth of the firm. A parallel shift in the term structure of interest rates would reduce the value of an S&L's long-term mortgages more than it would reduce the value of its short-term liabilities. Net worth would be reduced or possibly turn negative. Purchasing a cap—an asset on the balance sheet—would offset loss of net worth

to some extent because it would gain value as interest rates rise. See Spahr, Luytjes, and Edwards (1988) for a good exposition of this application of caps and how they hedge interest-rate risk.

[32]Commercial banks underwriting debt for highly leveraged financings often require their floating-rate borrowers to buy caps for a portion of the debt. This hedging requirement may be stipulated in the loan covenant. See Richardson (1989): 12.

[33]See *Moody's Special Report* (1989).

[34]This method assumes that the interest rate follows a random walk with no "drift" (that is, deterministic trend movements). Changes in the interest rate from period to period are assumed to be normally distributed with constant variance (or standard deviation), implying that the statistical distribution of interest-rate movements may be completely characterized by only its mean and variance.

[35]These percentages are based on the properties of the normal distribution, which is assumed to describe interest-rate movements.

[36]Arak, Goodman, and Rones (1986): 452.

[37]Cap valuation can be formulated as a kind of compound option problem. See Geske (1977) to appreciate the complexities involved in valuing securities that are composed of sequences of options.

[38]In a discrete time model the expected value is a weighted average of all possible payoffs, each payoff multiplied by the probability of its occurring.

[39]According to Moody's study, the lowest investment-grade bonds, rated Baa (or BBB by Standard and Poor's), had average annual default rates over two-year horizons of 0.25 percent. A Standard and Poor's BBB-rated investment bank was reportedly at a disadvantage in writing caps compared to stronger writers. See Shirreff (1986): 34.

The 0.13 default rate used in the example was chosen to reflect the lower risk of default on a cap relative to a bond.

[40]The Monte Carlo simulations used a variance reduction technique called the method of antithetic variates (see Boyle [1977]). The total number of realizations was in fact 200,000 for each simulation, though only a fourth of that number came from independent draws from the random number generator. See Abken (forthcoming) for more details.

References

Abken, Peter A. "An Introduction to Portfolio Insurance." Federal Reserve Bank of Atlanta *Economic Review* 72 (November/December 1987): 2-25.

__________. "Valuing Default-Risky Interest Rate Caps: A Monte Carlo Approach." Federal Reserve Bank of Atlanta Working Paper (forthcoming).

Arak, Marcelle, Laurie S. Goodman, and Arthur Rones. "Credit Lines for New Instruments: Swaps, Over-the-Counter Options, Forwards and Floor-Ceiling Agreements." In *Proceedings of a Conference on Bank Structure and Competition.* Federal Reserve Bank of Chicago (May 1986): 437-56.

Black, Fischer. "The Pricing of Commodity Contracts." *Journal of Financial Economics* 3 (January/March 1976): 167-79.

Boyle, Phelim P. "Options: A Monte Carlo Approach." *Journal of Financial Economics* 4 (May 1977): 323-38.

Boyle, Phelim P., and Stuart M. Turnbull. "Pricing and Hedging Capped Options." *Journal of Futures Markets* 9 (February 1989): 41-54.

Brown, Keith C., and Donald J. Smith. "Recent Innovations in Interest Rate Risk Management and the Reintermediation of Commercial Banking." *Financial Management* 17 (Winter 1988): 45-58.

"Caps and Floors." *The Banker* (February 1989): 9.

Chicago Mercantile Exchange. *Market Perspectives.* Various issues.

Commins, Kevin. "Managing Interest Rate Risk." *Intermarket* (May 1987): 28-34.

Cox, John C., and Mark Rubinstein. *Options Markets.* Englewood Cliffs, N.J.: Prentice-Hall, 1985.

Feinstein, Steven P. "Forecasting Stock-Market Volatility Using Options on Index Futures." Federal Reserve Bank of Atlanta *Economic Review* 74 (May/June 1989): 12-30.

Feinstein, Steven P., and William N. Goetzmann. "The Effect of the 'Triple Witching Hour' on Stock Market Volatility." Federal Reserve Bank of Atlanta *Economic Review* 73 (September/October 1988): 2-18.

Geske, Robert. "The Valuation of Corporate Liabilities as Compound Options." *Journal of Financial and Quantitative Analysis* 12 (1977): 541-52.

Henderson, Schuyler K. "Securitizing Swaps." *International Financial Law Review* (September 1986): 31-34.

Jarrow, Robert A., and Andrew Rudd. *Option Pricing.* Homewood, Ill.: Richard D. Irwin, Inc., 1983.

Johnson, Herb, and René Stulz. "The Pricing of Options with Default Risk." *Journal of Finance* 42 (June 1987): 267-80.

Kawaller, Ira G., Paul D. Koch, and Timothy W. Koch. "The Relationship between the S&P 500 Index and S&P 500 Index Futures Prices." Federal Reserve Bank of Atlanta *Economic Review* 73 (May/June 1988): 2-10.

Kuprianov, Anatoli. "Short-Term Interest Rate Futures." Federal Reserve Bank of Richmond *Economic Review* (September/October 1986): 12-26.

LeGrand, Jean E., and John P. Fertakis. "Interest Rate Caps: Keeping the Lid on Future Rate Hikes." *Journal of Accountancy* (May 1986): 130-36.

Mattu, Ravi. "Hedging Floating Rate Liabilities: Locks, Caps and Floors." Chicago Mercantile Exchange Strategy Paper, 1986.

Moody's Special Report. "Historical Default Rates of Corporate Bond Issuers, 1970-1988." July 1989.

Newbery, David M.G., and Joseph E. Stiglitz. *The Theory of Commodity Price Stabilization: A Study in the Economics of Risk.* New York: Oxford University Press, 1981.

Richardson, Portia. "Put on Your Thinking Cap." *Intermarket* (March 1989): 10-13.

Schaefer, Stephen M., and Eduardo S. Schwartz. "A Two-Factor Model of the Term Structure: An Approximate Solution." *Journal of Financial and Quantitative Analysis* 19 (December 1984): 413-24.

Shirreff, David. "Caps and Options: The Dangerous New Protection Racket." *Euromoney* (March 1986): 26-40.

Smith, Clifford W., and René M. Stulz. "The Determinants of Firms' Hedging Policies." *Journal of Financial and Quantitative Analysis* 20 (December 1985): 391-405.

Spahr, Ronald W., Jan E. Luytjes, and Donald G. Edwards. "The Impact of the Uses of Caps as Deposit Hedges for Financial Institutions." *Issues in Bank Regulation* (Summer 1988): 17-23.

Sutherland, L. Frederick. "Squeezing Cash: How to Make an LBO Work." *Corporate Cashflow* (June 1988): 47-50.

Tompkins, Robert. "The A-Z of Caps." *Risk* 2 (March 1989): 21-23, 41.

Wall, Larry D. "Alternative Explanations of Interest Rate Swaps: A Theoretical and Empirical Analysis." *Financial Management* (forthcoming, 1989).

Wall, Larry D., and John J. Pringle. "Interest Rate Swaps: A Review of the Issues." Federal Reserve Bank of Atlanta *Economic Review* 73 (November/December 1988): 22-37.

Federal Reserve Bank of Atlanta
1989 Annual Report

The Atlanta Fed's 1989 Annual Report will be available in February 1990. In addition to reviewing the Bank's activities during 1989, the report will include selections from its 75th anniversary publication, *A History of the Federal Reserve Bank of Atlanta, 1914 -1989.* It will also contain a statement of condition, a statement of earnings and expenses, and a statistical summary of operations, along with a list of directors and officers who served during the year.

For copies of the publication, please fill out the coupon below and mail to: Public Information Department, Federal Reserve Bank of Atlanta, 104 Marietta Street, N.W., Atlanta, Georgia 30303-2713, or call 404/521-8788.

Please send me ______________ copy (copies) of the Atlanta Fed's 1989 Annual Report.

Name ____

Address ____

____

City ______________________ State __________ ZIP ______________

Financial Asset Pricing Theory: A Review of Recent Developments

Ellis W. Tallman

In this article, an Atlanta Fed economist reviews research on financial asset pricing with a special focus on the links between asset pricing and the real economy. Surveying the capital asset pricing model (CAPM), the consumption-based CAPM, and the more recent arbitrage pricing theory, he concludes that ongoing theoretical and empirical developments point toward future research that can link real economic factors to asset pricing behavior.

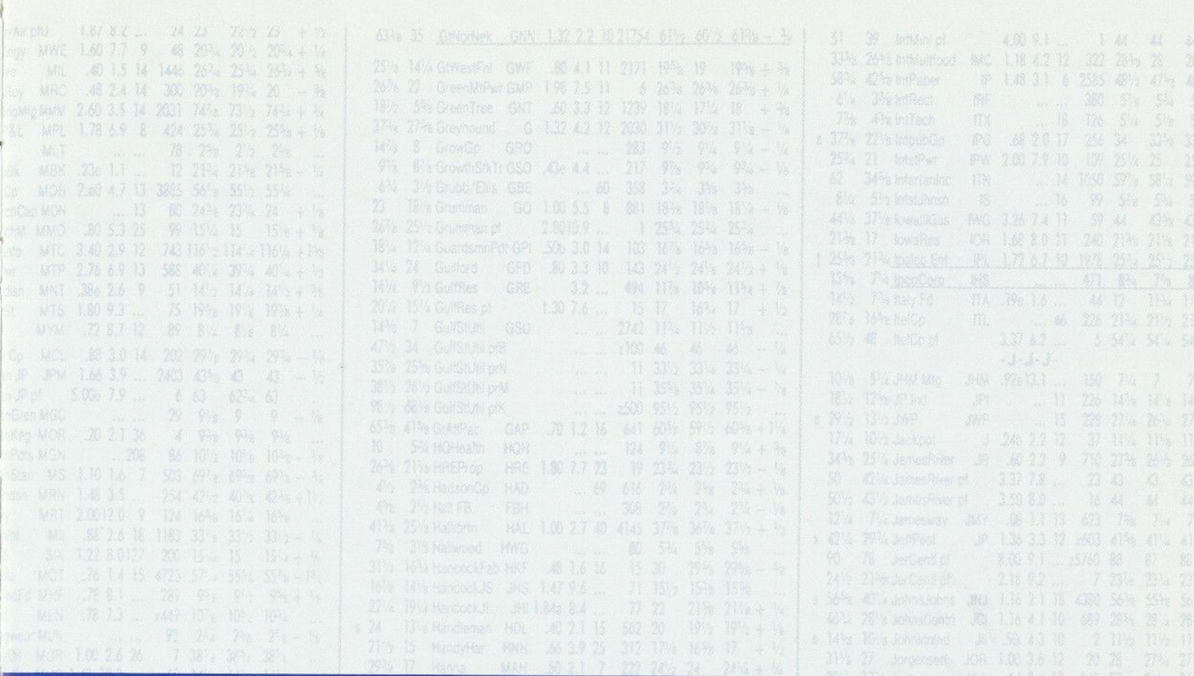
Financial asset pricing theories have developed primarily over the past 30 years. Scholars have made great strides during the latter half of the period in the analysis of newer, more dynamic models. This article surveys recent developments in theoretical research and the state of relevant empirical evidence. It concludes that financial asset pricing research remains open for additional study, yet the current body of knowledge presents a coherent framework for analyzing asset pricing in a rational economic setting.

Modern capital market theory studies the determination of asset prices. In a basic model of stock valuation, asset prices reflect the present discounted value of the projected future dividend payments to the stockholder. When new information about a firm's prospects becomes public, expectations about future cash flows or the risk-adjusted discount rate of a given stock change. In most models of asset pricing, investors, lenders, and other economic "agents" are assumed to be rational, using new information to adjust their valuation of the given firm's stock. Competitive market forces then bid the asset price up or down to its new equilibrium price.

This framework provides an intuitive link between the asset markets and measures of real economic behavior. Such a model suggests that the aggregate stock market value reflects expectations of the present discounted value of cash flows from the future performance of the economy. Despite the intuitive appeal of a correlation between performance of the stock market and the real economy, insufficient empirical evidence exists to support this relationship.

For example, from October 2 to October 23, 1987, nearly 30 percent of perceived asset value in the stock market, as measured by the Standard and Poor's 500, was lost. In contrast, the real economy grew and continued to grow throughout 1987 and 1988 and into 1989. In longer perspective, on the other hand, the overall growth rate of real gross national product (GNP) averaged 4.2 percent from 1983 through 1988, and the stock market boom has coincided, although not without volatility. The sometimes anomalous behavior of the stock market vis-a-vis real GNP growth indicates that the forces that drive asset prices are still largely a mystery, especially with regard to the relationship between real economic performance and asset markets.

The author is an economist in the macropolicy section of the Atlanta Fed's Research Department. He thanks Curt Hunter and Frank King for valuable comments.



To respond to a need for an update on the existing research on asset pricing, this article offers a general introduction to and historical survey of modern financial asset pricing theories; other surveys of asset pricing models have focused only on a specific subset of them. This paper concentrates on the recent developments in asset pricing models under uncertainty in the 17 years since an early survey of the groundbreaking research by Michael C. Jensen (1972). Its primary purpose is to show that such theories have provided useful advances in achieving a better understanding of equity market pricing.[1]

Presented first is a brief summary of the traditional capital asset pricing model (CAPM) and the groundbreaking work that developed from it. The survey then traces the evolution of the initial work on the CAPM to asset pricing in dynamic models, some of which encompass both the disciplines of finance and macroeconomics. While not exhaustive, the article concludes with suggestions for further work on asset pricing that may be especially useful to those involved in economic policy-making or in the application of asset valuation research.

Policymakers have a particular interest in better understanding linkages between real economic performance and asset pricing behavior. Financial data such as stock market prices are observed much more often than are most real economic data like industrial production and gross national product, which are available only monthly or quarterly. Such financial statistics may provide preliminary insights into the condition of real economic prospects, thus enhancing the policymaker's information set. Also, specific information on risk sources in the real economy can help policymakers promote a stable environment for financial markets and foster more efficient allocation of capital in the economy. Investment managers may also find the isolation of macroeconomic sources of asset risk a useful method of assessing portfolio risk and a criterion for portfolio formulation.

Valuation Theory, Mean Variance Efficiency, and the Traditional CAPM

As long as stock markets have existed, prognosticators have tried to predict future movements in equity values. Forecasts often lacked a strong theoretical and analytical framework, though. Early attempts to examine their effectiveness, notably by Alfred Cowles III (1933), suggested that they offered no perceptible advantages to investors. More formal and scientific analysis of asset price behavior began later.

Harry Markowitz (1952) provided the source for modern portfolio theory. His research is often cited as the seminal work of modern finance from which evolved the early impetus to describe the equilibrium relationship between assets and risk. The resulting asset pricing models, particularly the traditional capital asset pricing model, rely on the return to a "market" portfolio, consisting of a weighted average of all assets held, as the benchmark from which one can assess asset prices relative to the market. The CAPM provides a useful simplification and focus for asset pricing theory because it produces an interpretable risk measure for a risk-return relationship and parsimoniously summarizes a great deal of information in a single variable.

A weakness of the early equilibrium asset pricing models, however, is their lack of any formal linkage between real economic performance and the behavior of asset prices. Recessions—or, more generally, changes in future prospects for the economy—affect firm valuation by altering expected cash flows and the relevant discount rate, but the simple CAPM offers no direct method to incorporate fluctuations in economic conditions over time into the asset pricing process. The model ignores this issue. Recent theoretical advances in dynamic models of asset pricing as well as an accumulation of evidence in conflict with the simple CAPM have made that basic formulation less central as an equilibrium model of asset pricing.[2] Yet, the simple CAPM provides a major part of the underlying intuition for these models.

Modern financial theory has its underpinnings in the application of scientific methods to basic finance questions. Hundreds of different assets trade on the stock exchange, and each asset has characteristics, such as firm size, location, industry, and age, that distinguish it from other assets. The simultaneous analysis of so many characteristics is not feasible in a scientific realm. To focus attention on the most important traits, financial models simplify the problem by limiting the number of variables.

One of the earliest models of asset pricing, briefly mentioned above, analyzes the valuation of a single stock as a function of the flow of future dividends discounted by the relevant risk-related discount rate. The model, presented by John B. Williams (1931), is outlined below:

$$P_{i,0} = \sum_{t=1}^{\infty} \left(\frac{1}{1+k_i}\right)^t (d_{i,t}),$$

where $P_{i,0}$ is the price of asset i at period 0, $d_{i,t}$ is the dividend per share of common stock of firm i from the end of month $t-1$ to the end of month t, and k_i is the risk-related rate of discount for firm i.

Eugene F. Fama and Merton H. Miller (1972) show that—given a number of simplifying assumptions—firm valuations derived from discounted firm cash flows, the stream of dividends, or the firm's earnings produce Williams's result. Williams's model, in a deterministic world without uncertainty, thus offers a framework to guide analysis. This valuation model, altered for uncertainty, suggests that asset prices vary upon the release of new information regarding a firm's prospects.

Before the announcement of the return on a stock, a large degree of uncertainty exists regarding the actual outcome, that is, the return ex post.[3] In modern financial theory, stock returns are viewed as random variables, and a probability distribution is associated with them. For most applications, stock return distribution has been assumed to be approximated by the normal distribution, which is fully summarized by two parameters: the mean, which is the measure of central tendency, and the variance, the measure of dispersion around the central tendency.[4]

Markowitz presented a model of investor portfolio selection under uncertainty in which investors choose asset portfolios on the basis of asset return and variance in a single period. Portfolio optimization involves the trade-off between reward (expected return) and greater risk. Investors prefer assets with higher means of expected returns but lower return variances, or less return variability. Thus, investors, assumed to be risk-averse, want to balance risk and return in their portfolio choice.

A further insight of portfolio theory is that the addition of a security adds to a portfolio's risk mainly by the contribution of its variability to the variability of return from the entire portfolio—its "covariance" with the return stream of other portfolio assets. In the limit, as one increases the number of individual assets in a portfolio, this covariance risk is the dominant component of financial asset portfolios' variance.

This insight regarding covariation suggests that a collection of assets can offer a lower level of return variability than individual assets held separately. Consider two assets whose returns are perfectly negatively correlated. (The percentage increases in returns to one asset occur simultaneously with equal but opposite movements in another asset's return.) A portfolio of these two assets would carry a fixed return and no risk, or variability of return. Although such neat covariation properties of assets do not often occur, the portfolio helps illustrate the advantages of diversification.

A primary outgrowth from the Markowitz work was the general equilibrium models of asset pricing credited most often to William F. Sharpe (1964), John Lintner (1965), and Jan Mossin (1966).[5] These models assume the existence of an asset that is both free of default risk and offers a fixed one-period return.[6] In equilibrium, all assets are held, and the market portfolio—as defined earlier, the portfolio that represents the return on every asset weighted by its proportion in the total value of all assets combined—comprises entirely risky assets. James Tobin (1958) shows that in Markowitz's environment, all individuals hold assets in only two types of portfolios: the riskless asset and the market portfolio.

The traditional capital asset pricing model, with fixed covariance between returns on an asset and returns on the market portfolio, shows that the risk premium of an asset (that is, the difference between the return on a risky asset and the return on a risk-free asset) is determined by movements in the market portfolio's expected premium. For the purpose of this article, the following equations present the main implications of the model:

$$E_i = R_F + \beta_i (E_M - R_F)$$

and

$$\beta_i = \frac{Cov(R_i, R_M)}{\sigma^2(R_M)},$$

where E_i is the expected return to asset *i*, R_F is the rate of interest on the riskless asset, E_M is the expected return on the market portfolio, β_i is the degree to which asset *i*'s return varies with the market's return (discussed below), $\sigma^2(R_M)$ is the variance of the return to market portfolio, and $Cov(R_i, R_M)$ is the covariation between the return to asset *i* and the return on the market.

The main argument of the model relies on the intuition that investors are rational and will undertake risk only to the extent that they are compensated. If, simply through diversification, risk can be removed from a portfolio, no one should be compensated for holding risk that can be avoided—"diversifiable" risk. While diversification can be achieved by holding assets that should have low or negative covariation, Markowitz's portfolio theory result suggests that holding a large number of assets also results in diversification.

Diversifiable risk should not be related to a risk premium. If a firm experiences a period of poor management or suffers a labor strike, the asset returns may be negative. But these sources of risk are company- or firm-specific, and an investor can reduce risk by investing a proportion of wealth in other firms. In contrast, certain factors like wars or the oil price shocks of the 1970s affect the entire economy and, as a result, future returns to the market portfolio. Such *nondiversifiable risk*—that is, risk related to covariation of an asset's return with the return to the market portfolio—will therefore be related to a risk premium since an investor will require an incentive to hold a risky group of assets.

For an individual asset *i*, the expected return equals the riskless rate of interest plus the product of the market risk premium and the relevant risk measure, β_i, commonly referred to as the risk of covariation with the market, or "beta" risk. An implication of the model is that assets are priced relative to their sensitivity to the market portfolio returns. A portfolio in which β equals one results in the same expected return as the market portfolio. Portfolios in which β is less than one are referred to as defensive since they should fluctuate relatively less than the market but will also have a lower expected return. In contrast, portfolios with β greater than one are deemed aggressive, in that their expected returns are greater than the market portfolio's, but they incur relatively greater return volatility. Thus, according to the CAPM only nondiversifiable, or systematic, risk is relevant for asset pricing: consequently, the expected return on any asset is a linear function of the asset's β. One may interpret β as a sensitivity measure, gauging the reaction of the return on asset *i* to a movement in the market return. Any returns that are significantly greater or less than predicted by an asset's beta-risk measure are called *abnormal returns.*

Testing the Capital Asset Pricing Model. Despite CAPM's simplicity and the measurability of its variables through market price data, empirical evidence has not supported this approach. Two important empirical tests of the CAPM—Fischer Black, Jensen, and Myron S. Scholes (1972) and Fama and James D. MacBeth (1973)—find evidence that conflicts with the predictions of the simple Sharpe-Lintner pricing model introduced earlier. Black, Jensen, and Scholes find evidence that low beta risk stocks or portfolios have positive abnormal returns, and high beta risk stocks or portfolios have negative abnormal returns.[7]

Fama and MacBeth included other measures as additional explanatory factors in an asset pricing regression to examine the CAPM's sensitivity to variables that theory suggests should be unimportant. The added measures—beta squared and the average of the residual variance—might also indicate possible nonlinearities in the risk-return relationship. Although over the entire sample period their two additional variables show no significant systematic relationship to priced risk, in certain time periods these measures were associated with statistically significant risk premia—that is, the return which asset holders must be paid in order to induce them to accept an asset with nondiversifiable risk. Fama and MacBeth concluded that these variables serve as proxies for relevant underlying risk measures.[8] However, their results suggest that a positive trade-off generally exists between risk, as measured by beta, and return.

In a recent paper that reexamines and extends the Fama and MacBeth estimates, Seha M. Tinic and Richard R. West (1986) showed both significant departures from the linear risk-return trade-off predicted by the traditional CAPM and significant nonlinearities not captured by the model. As a result, the researchers conclude that results of existing empirical research on the traditional CAPM is suspect.

Much of the empirical literature that uses the CAPM applies beta as a risk measure in order to adjust asset returns for their degree of riskiness, prior to the examination of the impact of an event.[9] For example, beta has been used to adjust returns for risk in numerous event studies dealing with finance issues like judging mutual fund performance or the price effects of stock splits and public tender offer announcements.

After rigorous investigation, researchers have uncovered a number of anomalies that undermine the capital asset pricing model. Two of the more well-known inconsistencies, known initially as the *January effect* and the *small firm effect,* have been particularly damaging. Researchers have found that the returns to small firms generally outperform those of larger firms after adjustment by beta risk measures. In other research, stocks have shown an abnormally high excess return after risk adjustment in the month of January. A study by Donald B. Keim (1983) has shown that the small firm effect and the January effect are related; that is, smaller firms outperform larger firms in January. These results contradict implications of the CAPM. They have fueled criticisms of its framework not only as an equilibrium model of asset pricing but also as a useful framework for risk adjustment in other applications.

"Despite CAPM's simplicity and the measurability of its variables through market price data, empirical evidence has not supported this approach."

Recent work by Jay R. Ritter and Navin Chopra (1989), however, suggests that the capital asset pricing model's risk-return relationship is more robust when portfolios weight the individual assets by their proportion of total market value in contrast to the standard practice of weighting assets in a portfolio equally. In those circumstances, small firm effects are deemphasized and the relationship shows no January seasonal effects. If nothing else, then, the existence of anomalies has stimulated additional research and evidence on the traditional CAPM.

Aside from these empirical shortcomings, the CAPM is essentially not dynamic. Though the model involves numerous simplifying conditions, for this article the most relevant assumption is that all investors maximize the utility of terminal wealth. In other words, notwithstanding inevitable uncertainty, future investment and consumption opportunities are completely captured by the certain mean and variance of the probability distribution of the asset returns. The CAPM is hence a one-period model that cannot encompass issues like economic fluctuations and their effects on asset pricing. Fama (1970) has shown that the CAPM's assumption about maximizing terminal wealth can be extended to multiple periods as long as future consumption and investment decisions are determined outside the model. However, the implication of Fama's work is that investors have perfect foresight with regard to future market conditions; intertemporal factors like changes in economic performance are assumed to have already been anticipated and thus should not affect the investment decision. In this sense, the model is static, although empirical tests estimate the model as if its restrictions held over time. Consequently, aside from predicting the return to the market portfolio (a difficult task at best), the CAPM as interpreted by Fama does not present a method to estimate any dynamics.

"Recent advances in the sophistication of financial asset pricing theory have moved to models that address intertemporal variations in opportunities over the business cycle."

Another controversy in CAPM-based asset pricing research is whether a truly riskless rate of return actually exists. Black (1972), in fact, attacks this question by offering a CAPM without the riskless rate. Generally, though, in empirical tests, the riskless rate of return, R_F, is proxied by the return on a Treasury bill with one month to maturity. The return to the market portfolio, R_M, is approximated by some equity index, usually the Standard and Poor's 500.

Richard Roll (1977) criticizes empirical examinations of the traditional CAPM in what is now referred to as the "Roll critique." His primary concern is measurement of the market portfolio in CAPM tests. Roll argues that an unambiguous test of the model cannot be performed with the typical proxies for market rate of return because the true market portfolio has to include all individual assets. The argument suggests that inferences about the model may be sensitive to the composition of the market proxy, and any demonstrated sensitivity to various reasonable proxies for the market will reduce the testability of the model. Robert F. Stambaugh (1982), responding to this criticism, has shown that, although tests are sensitive to the selection of assets, inferences about the CAPM are insensitive to the use of several different proxies for the market portfolio, suggesting that the CAPM may be less sensitive to the Roll critique than the argument implies. Still, Roll's analysis has contributed to deemphasizing the model in more recent research.

Intertemporal CAPM and the Consumption CAPM

The ambiguous empirical support for the traditional CAPM as well as dissatisfaction with the restrictiveness of some of its assumptions has led researchers toward models that relax some CAPM assumptions. Recent advances in the sophistication of financial asset pricing theory have moved to models that address intertemporal variations in opportunities over the business cycle. Empirical evidence suggests that the extension of the CAPM to account for intertemporal change is useful. For instance, Katherine Schipper and Rex Thompson (1981) demonstrated that equity assets may be used to hedge against changes in consumption and investment opportunities related to unanticipated shifts in consumption, GNP, and the price level, which serve as proxies for general conditions.[10]

Robert C. Merton (1973) earlier developed a dynamic asset pricing model, drawing from the initial insights of the mean-variance CAPM but extending the framework to incorporate intertemporal uncertainty. Merton's pricing equation describes a framework that holds in the presence of a business cycle. The model allows investment and consumption opportunities to fluctuate over time so that the economy's condition is linked directly with asset price behavior. John B. Long, Jr., (1974) introduced an alternative dynamic model that specifies relevant variables, known as "state" variables because they indicate the state of the economy, useful for the pricing equation. These state variables represent external factors, such as the stock of physical capital, that determine current investment opportunities.[11] Long's model suggests that the term structure of interest rates—that is, interest rates on equally risky debt of successively distant maturities—is a key element in the pricing of equity assets, and recent empirical work has supported this intuition.[12]

The general model in Merton (1973) also involves a vector of state variables, which represents the number (S) of relevant variables needed to describe the condition of the economy. The state vector indicates whether the economy is in a recessionary or expansionary stage of the business cycle and characterizes uncertainty in a model economy.

The model solution, in the general case, implies that there will be $S + 2$ number of portfolios in the equilibrium asset pricing equation. The additional two portfolios are the market portfolio and the riskless asset.[13] The resulting pricing relationship would expand the simple CAPM to include S additional betas (or sensitivity measures, one for each state variable), and the premia related to each state variable sensitivity. The general model requires identification of state variables that may be unobservable, however. Thus, the model may not be directly empirically testable with existing econometric methods.

To give more interpretation to the model as well as to develop a potential route of inquiry, Merton assumes that the interest-rate movements of the riskless asset are the sole state variable sufficient to describe the investment opportunity set. The restricted model provides a tractable result in which the equilibrium asset pricing model involves three funds: the market portfolio, the riskless asset, and a portfolio of assets negatively correlated with movements in the riskless asset.

This simplified model presents the intuition of intertemporal uncertainty more directly than does the general specification. Investors are compensated for holding both market risk, just as in the static CAPM, and the risk of unfavorable movements in investment opportunities as conveyed by the riskless interest rate. In periods of poor economic opportunities, investors would like to have assets that offer large returns. In fact, an investor may hedge against aggregate intertemporal risk by holding a risky asset that has an expected return less than the riskless asset if the risky asset pays off a high return when the return to the riskless asset is low.

Although the restricted model in Merton's work provides insights into the forms of risk presented by shifts in aggregate economic opportunities, the three-fund result remains a special case. Testing Merton's general model requires identifying and counting the state variables.

"[An alternative] model allows investment and consumption opportunities to fluctuate over time so that the economy's condition is linked directly with asset price behavior."

Unfortunately, theory makes no unambiguous predictions about their number or identity.[14] However, the model motivates investigation of additional variables as measures of intertemporal risk and offers a general structure for empirical analysis.

The Macroeconomic Link. In an elegant and influential paper, Douglas T. Breeden (1979) provides a key link between macroeconomic growth models and financial models of asset pricing. This connection makes possible an analysis of asset price determination in a model economy that fluctuates over time. Breeden's construct, which is consistent with Merton's, shows that the growth rate of consumption is a sufficient statistic for the state of the economy; in other words, the S state variables need not be identified for asset pricing. The resulting relationship is an equilibrium asset pricing model that uses the growth rate of (real per capita) consumption as the benchmark return from which all other assets are priced.[15] Thus, covariation with consumption growth is the single relevant measure of risk. Breeden argues that aggregate consumption should be a better proxy for the desired measure of consumption than the return to a market proxy is for the return to the market portfolio. The model is commonly referred to as the consumption CAPM (CCAPM).[16]

The equilibrium relationship is:

$$E_i - R_F = \beta_{ci} (E_c - R_F),$$

where E_i is the expected return to asset i, R_F is the riskless rate of return, E_c is the expected growth rate of real per capita consumption, and β_{ci} is the measure of the covariance of an asset's return with the growth rate of consumption.

"Its simplicity as well as its derivation from a dynamic model has made the consumption CAPM an attractive method of asset pricing."

By using consumption as the benchmark, the model is implicitly concerned with fluctuations in both consumption and investment opportunities. The business-cycle behavior of consumption, therefore, directly affects the pricing of assets. The intuition behind this relationship is that the marginal utility, or marginal contribution to valuation from an extra unit of consumption, is low in a time period that has high consumption. If consumption fluctuates, a consumer would prefer assets that will help reallocate consumption across states to those in which consumption is low. As a result, an asset return that covaries negatively with consumption growth should help smooth consumption and be associated with a negative risk premium. On the other hand, asset returns that covary positively with consumption are associated in this model with a positive risk premium. Hedging behavior on the part of the investor results from the incorporation of intertemporal uncertainty into the model. Persistent differences in average yields to a selection of assets can be explained, therefore, by the insurance that particular assets provide against certain states.

Its simplicity as well as its derivation from a dynamic model has made the consumption CAPM an attractive method of asset pricing. As with the traditional CAPM, the equilibrium relationship of the consumption model requires estimating only one parameter to evaluate the risk characteristics of an asset or portfolio. Consumption data are also readily available on a monthly basis, so the model can be tested relatively easily. It has been tested often.

A theoretical criticism by Bradford Cornell (1981) suggests that the consumption model is not free of the restrictions implied by Merton's intertemporal model; direct estimation still requires the identification of state variables. As a result, the conditional distribution of consumption betas is random. Although Cornell notes that this situation may be resolved with empirical evidence, the theory implies that distribution of the consumption betas relies upon the properties of the state variables.[17]

Despite Cornell's criticism, some empirical research has been done on the adequacy of the consumption CAPM. A recent study by Gregory N. Mankiw and Matthew D. Shapiro (1986), using quarterly data from 1959 to 1982, shows that the traditional CAPM outperforms the consumption CAPM. Based on a large sample of equity returns, their test employs instrumental variables estimation methods.[18] Its results show that the expected real return has a significant linear relationship with the market beta but not with the consumption beta.

Simon Wheatley (1988) criticizes the inferences made from the Mankiw and Shapiro results because the instrumental variables estimations are widely different from those using the ordinary least squares regression technique, suggesting problems with the selected estimation strategy and weakening the resulting inferences.[19] Wheatley continues by estimating the cross-sectional adequacy of the consumption CAPM restrictions using 40 stock portfolios, Treasury bills, Treasury bonds, and corporate bonds as dependent variables. His tests found the CCAPM implications consistent with the data.[20]

To transcend Cornell's criticisms, Breeden, Michael R. Gibbons, and Robert Litzenberger (1989) used weaker empirical tests of the consumption model. Their paper examined the CCAPM relative to the traditional CAPM using 12 stock portfolios, a Treasury bill asset, a Treasury bond portfolio, a corporate bond portfolio, and a junk bond premium as the set of asset returns to be explained. Their consumption data are adjusted for measurement problems associated with reported consumption statistics. The primary problem with aggregate consumption data is that they are issued much less frequently than observations of stock returns. The CCAPM theory requires measurement of spot consumption growth rates, whereas actual consumption is measured over an interval. Given the data adjustments required, empirical evidence found in support of the model is necessary, but not sufficient, to accept it.

Nonetheless, Breeden, Gibbons, and Litzenberger find that the explanatory power of the consumption growth rate for the behavior of asset returns over time is significant. The results using the return to a portfolio of assets that has maximum correlation with consumption growth—the *maximum correlation portfolio*—are more comparable to linear regressions with the return to a market proxy, since they both use portfolio returns data as independent variables. Tests of linearity between consumption beta and expected returns reject the hypothesis that consumption β and expected returns are linearly related for the full period covered by the data. However, examination of the subperiods suggests that the source of the rejection is the period 1929-39, a unique time in the U.S. economy. In all other subperiods, the relationship cannot be rejected. Test results imply that neither the maximum correlation portfolio nor the market portfolio proxy has the lowest variance for a given mean return, that is, neither proxy is mean-variance efficient. Despite these rejections, the estimates for the risk premia related to consumption and to the market are quantitatively similar.

Macroeconomic studies that embody the consumption capital asset pricing model concentrate on the relationships between forecastable movements in asset returns and in consumption. Lars P. Hansen and Kenneth J. Singleton (1982, 1983) imposed strong assumptions to generate a closed-form solution that would test the predictability of asset returns and obtain estimates of the structural parameters of interest, namely the degree of risk aversion and the intertemporal discount factor.[21] Unfortunately, their empirical results suggest rejection of the model, in part because of problems associated with measuring consumption data. However, further work by Sanford J. Grossman, Angelo Melino, and Robert J. Shiller (1987), which explicitly accounts for the time averaging of consumption data, also failed to support the model.[22]

In sum, the empirical results for the consumption CAPM are mixed. The strong restrictions imposed by the macroeconomic model tests lead to rejection of the model and do not produce reasonable or useful estimates of the structural parameters. In contrast, recent evidence on the consumption CAPM as a relative asset pricing construct are more hopeful, suggesting some potential for its use in evaluating asset risk. However, more research is necessary to determine the robustness of the consumption beta measure as a risk gauge for assets; the initial tests are supportive of a linear relationship between consumption risk and asset premia but do not support mean-variance efficiency, a prediction of the model. Thus, further research on the consumption CAPM must be done before it can be widely applied.

Arbitrage Pricing Theory

An alternative to the traditional CAPM paradigm that has gained considerable attention is arbitrage pricing theory, developed by Stephen A. Ross (1976).[23] This model retains the distinction between diversifiable and nondiversifiable risk but imposes fewer restrictive assumptions in its derivation of asset returns than does the CAPM. For example, the traditional pricing model requires that returns follow the normal distribution, implying that knowledge of the mean and variance is sufficient to describe the entire distribution. The traditional CAPM relies on the return to the market portfolio as the benchmark variable that describes asset return behavior relative to it. In contrast, arbitrage pricing theory does not require normally distributed returns and suggests that a number of variables, known as factors (risk sources), describe asset returns.

The derivation of arbitrage pricing theory requires two major assumptions. First, agents are assumed to believe that some identifiable set of factors generates the variability of all asset returns and that their relationship is consistent across the range of variables. The second assumption is that no opportunities are available for riskless arbitrage (that is, no unlimited profits given no net investment). The hypothesized linear factor model is:

$$R_i = E_i + \sum_{j=1}^{k} b_{ij}\,\delta_j + \varepsilon_i,$$

where R_i is the uncertain return to asset *i*, E_i is the expected return to asset *i*, b_{ij} is the factor loading for asset *i* related to factor *j*, or asset *i*'s sensitivity to movements in factor *j*, δ_j is the factor *j* ($j = 1, \ldots, k$), and ε_i is the error term for asset *i*. In addition, the model assumes that the factors and error terms have a mean of zero. It does not make other assumptions about the distribution of the factors or error terms aside from requiring that the covariance between the error terms, ε_i and ε_j, is zero.

In the derivation of the traditional capital asset pricing model, the "market" model that relates all asset returns to movements in the market return (prior to the pricing equation) follows directly from the assumption that returns are jointly normally distributed. In arbitrage pricing theory, however, the linear factor model is an assumption, although the idea that a set of forces determines the movements of all asset returns is compelling.

Exact arbitrage pricing implies the following asset pricing relationship:

$$E_i = \lambda_0 + \sum_{j=1}^{k} \lambda_j\, b_{ij},$$

where λ_0 is riskless or zero *b* return and λ_j represents risk premia related to factor *j*. A clear intuition underlies the equilibrium relation of arbitrage pricing theory. If the indicated factors truly generate the movements of all asset returns and if current asset prices allow no riskless arbitrage, it follows that expected returns are approximately linearly related to covariance between the asset returns and the factors. A main contribution of this theory is that it recognizes the importance of covariance risk in asset pricing as a result of the no-arbitrage assumption, which has strong theoretical appeal.

Arbitrage pricing theory is an attractive generalization of the traditional CAPM model's insight that covariance risk—risk that cannot be diversified away—underlies the pricing of assets.[24] The arbitrage pricing model provides a coherent structure, less restrictive than the CAPM, that allows for investigation of the sources of risk. The linear factor model framework, in addition, appears better able to account for the anomalies that conflict with the traditional model's predictions. In arbitrage pricing theory, the covariance is measured relative to the factors that determine the behavior of asset returns, whereas the CAPM gauges covariance only relative to the market return. Thus, finding a size factor or a seasonal factor that explains the CAPM anomalies would seem possible.

Arbitrage pricing theory has few underlying assumptions. It has been criticized, though, because its initial form has few rejectable hypotheses. Refuting the theory itself, which does not identify or limit the number of factors, is difficult.[25] Thus, tests of the arbitrage pricing theory are combined examinations of the pricing relationship and the appropriateness of the set of factors chosen.[26] Theoretical extensions and refinements by a number of researchers have provided the foundation for the substantial amount of empirical research that has been produced and the many works that are still in progress.[27]

Empirical methods developed to implement estimation of the factors and factor loadings in arbitrage pricing theory have involved two distinct approaches to the data: factor analytic techniques (or principal component analysis as in the work of Gregory Connor and Robert A. Korajczyk [1988]) and prespecification of the factors.[28] The former method employs the estimated covariance matrix of returns to determine the factor structure that underlies asset return behavior.[29] Estimates of the factors are determined in accordance with arbitrage pricing theory; that is, factors are estimated from the characteristics observed in the set of returns. The second technique attempts to identify factors without first examining the structure of returns. Instead, variables are chosen as needed by economic intuition that these factors affect asset pricing. The method uses the prespecified factors to estimate factor loadings and then tests to see if the loadings are associated with significant risk premia.[30]

This article will survey only the most recent of the large number of papers on arbitrage pricing theory.[31] Two recent works compare this theory with the traditional CAPM approach as models of asset pricing.[32] Bruce N. Lehmann and David M. Modest (1988) cannot reject the arbitrage-based construct when asset portfolios are formed on the basis of dividend yield or an asset's own return variance. Since CAPM research has found an anomaly with regard to dividend yield and asset pricing, the evidence can be viewed as supportive of the arbitrage pricing theory as an alternative.

On the other hand, the researchers reject the arbitrage pricing model when the portfolios are formed on the basis of firm size. Connor and Korajczyk (1988) show evidence consistent with Lehmann and Modest that a significant relationship exists between firm size and asset expected return that is not captured by the arbitrage pricing theory. However, Connor and Korajczyk demonstrate that the size effect is separate from a seasonal effect (for example, the January effect), which appears to be explained by the variation in the risk factors.[33]

K.C. Chan, Nai-fu Chen, and David A. Hsieh (1985) and Chen, Roll, and Ross (1986) employ prespecified factors in testing the arbitrage pricing theory predictions. In both works, the theory shows no significant anomalous behavior related to firm size.[34] Since the two empirical methodologies are quite different, the conflicting evidence suggests that more research is needed to investigate the issue.

In sum, arbitrage pricing theory represents a generalization of the CAPM intuition that covariance risk forms a basis for asset pricing. The arbitrage model makes few assumptions in its derivation and provides a method to investigate the underlying sources of asset risk. This theory has shortcomings, however, especially with regard to the number and the identity of the underlying risk factors and empirical testing. The factor structure has been estimated in different ways, although the techniques that estimate this structure from the estimated covariance matrix of asset returns seem most consistent with the theoretical model.

The current status of arbitrage pricing theory testing suggests its consistency with the data, although the model cannot explain the anomalous firm size effect. However, the capital asset pricing model cannot account for that anomaly, either. Thus, among static models, arbitrage pricing theory seems to be a viable alternative to the CAPM.

Dynamic Models

Although theoretical extensions of both paradigms have extended the basic model to an intertemporal realm in search of a productive way to link pervasive economic factors to equity market performance, both arbitrage pricing theory and the capital asset pricing model employ the fiction of a single period model.[35] The linkage between macroeconomic and financial markets has also been explored within a dynamic general equilibrium model, notably by William A. Brock (1980, 1982), and John C. Cox, Jonathan E. Ingersoll, Jr., and Ross (1985), in which asset prices are endogenous functions of underlying economic forces. However, the empirical implications of these models are not directly testable.

Brock provides a key link to understanding the relationship between the static and intertemporal models, and he emphasizes the degree of generality in the models that generate the testable implications. The Merton pricing model, for example, introduced intertemporal aspects to the traditional model, but Merton's is a partial equilibrium construct. The state variables that determine intertemporal asset price movements are not linked to the underlying sources of uncertainty in the economy. Brock, on the other hand, derives a general equilibrium model in which the state variables—here, technological factors underlying economic uncertainty—determine behavior of asset prices. The model provides one interpretation of technological shocks as the arbitrage pricing theory factors presented in Ross (1976), but it may also apply to Merton's model, or Long's (1974). In an example, Brock shows that the Sharpe-Lintner capital asset pricing model conforms to the case of one underlying technological shock. As a further exposition, the Brock model has the characteristic of consumption sufficiency—as in the Breeden model—so that the consumption CAPM model holds. Thus, Brock presents a framework in which all major asset pricing models may be derived and provides a unifying system to motivate research linking macroeconomic factors and asset price behavior.

Empirical research on arbitrage pricing theory that employs prespecified factors to estimate factor loadings relates closely to the macro-finance implications of the Brock as well as the Cox, Ingersoll, and Ross models. In fact, Chen, Roll, and Ross use these more general models on which to base their empirical investigation. As mentioned above, the method appears to stretch the arbitrage pricing theory's motivation of the determination of factor structure, yet they address the issue of the factors' identity. The motivation of linking the real economy with asset returns and the freedom to choose factors a priori has produced some stimulating empirical research.

Chan, Chen, and Hsieh (1985) investigated the firm size effect by prespecifying factors in a multifactor pricing equation, in which the factors are measures of economic and financial activity that may relate to asset pricing. They include the market portfolio, industrial production, two estimates of inflation, the change in the term structure, and the risk premium. The results show that the observed firm size effect has a strong relationship with the risk premium measure (the difference between the yield on low-rated long-term bonds and the yield to a portfolio of long-term government bonds). The variation in the risk premium reflects alterations in business conditions and, therefore, introduces an intertemporal feature into the empirical asset pricing model. These results imply that the firm size effect may be captured by the multifactor asset pricing model. Also, the firm size effect may be consistent with an efficient market in which small firms have higher expected returns because of higher risk that is not captured by the traditional CAPM model's risk measures.

Subsequent empirical work by Chen, Roll, and Ross investigated directly the role of economic forces in asset pricing, using similar economic variables as the prespecified factors. They found similar evidence that the multifactor model explains the pricing of a selection of asset portfolios formed on the basis of firm size. The authors shaped the selection of economic state variable proxies by choosing those that influence either cash flows to firms or the discount rate applied to asset cash flows in the simple stock valuation model.

The formulation of prespecified factors in these studies assumes that the chosen variables constitute the factor structure of asset returns, which is the underlying determinant of asset return time variation. The research does not present time-series evidence to suggest that these factors explain much of the time-series variability of asset returns. In factor analytic research, time-series explanatory power is evident in the method of identifying factors. In future research on economic factors and asset pricing, the time-series regressions of the prespecified factors will be useful indicators of whether the chosen factors are relevant.[36]

The underlying shocks (or factors) in Brock and Cox, Ingersoll, and Ross represent technological shocks that directly affect the productivity of the economy. Yet the main explanatory variables in these studies are financial measures, notably the term structure proxy and the risk premium proxy. These variables are at least partially determined by the true underlying factors, just as asset prices are. Although financial factors have significant implications for the pricing of risky assets and provide insights into the interrelationships of macroeconomic and financial markets, the results do not reveal the underlying sources of uncertainty.

The search for these underlying sources may seem futile. However, recent studies by David Alan Aschauer (1989a, b) suggest that government spending behavior, primarily changes in the public capital stock, may be one source of uncertainty which directly affects the aggregate productivity of the economy. In his 1989 paper, this researcher shows that the government stock of infrastructure capital—for example, roads, buildings, sewers, and so on—has a significant impact on the profitability of the aggregate economy. In other words, the public capital stock has a positive effect on the aggregate value of private firms. Thus, further work on asset pricing should investigate the effects of government policy, since these shocks seem most justifiably to be exogenous variables.[37] Research in this area may also provide economic policymakers with better information about the long-term effects of spending policies at all levels of government.

Conclusion

Recent financial models of asset pricing derive from the initial insights into the relative riskiness of different assets and the trade-off between risk and expected return provided by the traditional capital asset pricing model. The various models surveyed here provide a coherent framework in which to analyze asset returns. Over time, the CAPM has been useful for portfolio evaluation and for extending scientific analysis of financial markets. It continues to be used as a method to evaluate the risk of assets or portfolios. The model requires only one estimable parameter per firm or portfolio. However, the CAPM's static framework, the difficulty of predicting the return to the market proxy, evidence in conflict with its implications, and theoretical advances in financial modeling have combined to shift attention away from the simple CAPM as an equilibrium model of asset pricing. The consumption-based asset pricing model (CCAPM) presents a dynamic construct in which a single estimable parameter measures asset risk. This model has met criticism on matters of data measurement and inconclusive empirical evidence of its usefulness. Recent work nonetheless presents some support for further research in this area.

The intertemporal CAPM and arbitrage pricing theory, though clearly different models, suffer similar empirical difficulties. Neither provides insights into the identity of the multiple sources of asset risk. For arbitrage pricing theory, empirical applications using factor analysis cannot interpret risk sources. However, the two models provide a motivation for investigating multiple sources of asset risk.

Ongoing theoretical developments point toward future research that can link economic factors to asset pricing behavior. Such research should interest investors and especially policymakers, who may gain insight into the effects of alternative economic policies. For the policymaker, an appreciation of the risk sources in the economy can aid formulation of policy by linking information provided by equity market behavior to real economic performance. In addition, further research may uncover elements of economic policy as sources of macroeconomic uncertainty and provide policymakers with an improved set of policy guides.

Notes

[1] In fact, as evidence of their influence, financial asset pricing theories have penetrated Wall Street and are currently being used in the design of mutual fund portfolios.

[2] The traditional capital asset pricing model remains, however, a useful method to analyze asset characteristics with few measures.

[3] This uncertainty is larger in some periods than in others.

[4] Despite evidence that stock returns appear leptokurtic ("fat-tailed," or having a higher probability than a normal distribution of observing extreme values), the normal distribution remains the most common approximation of the distribution of stock returns.

[5] See also Treynor (1961).

[6] The traditional CAPM involves several additional assumptions that may be found in Sharpe (1985) or other finance textbooks.

[7] Black (1972) presents a model without a riskless asset that offers predictions that are more consistent with the Black, Jensen, and Scholes results.

[8]Levy (1978) suggests that if individuals are not well diversified their own variance of return should be relevant for asset pricing.

[9]For an intuitive discussion of event study methodology see Hunter and Walker (1988).

[10]The results, however, do not represent a test of any specific model.

[11]One may interpret these state variables to be virtually anything: the weather, oil supply shocks, measures of government policy, and so on. As will be discussed further, the identity of such variables is subject to considerable debate.

[12]See results in Chan, Chen, and Hsieh (1985); Chen, Roll, and Ross (1986); and McElroy and Burmeister (1988).

[13]The model has also been referred to as the multibeta CAPM, in which an estimated parameter is associated with each state variable.

[14]The problem, however, is not unique to Merton's model, as the discussion of arbitrage pricing theory indicates.

[15]One could also use the asset portfolio that has maximum correlation with the growth rate in consumption as the benchmark.

[16]See also Rubenstein (1976) and Breeden and Litzenberger (1978).

[17]See Cornell (1981) for technical elements and the complete argument. Bergman (1985) criticizes the assumption of time-separable preferences in the derivation of the CCAPM. This assumption implies that past decisions on consumption do not affect today's choices. Without the assumption, the Merton ICAPM still holds, but it can no longer be collapsed into the CCAPM. Despite this problem, most macroeconomic models assume that there are time-separable preferences and that the CCAPM holds.

[18]Instrumental variables estimation methods use variables correlated with the regressors but unrelated to the errors in an attempt to reduce the potential correlation between regression variables and the residual error.

[19]Wheatley suggests that either the instruments are weakly related to the underlying variables of interest or that the underlying variables are collinear. In either case, the results in Mankiw and Shapiro (1986) are suspect.

[20]Although the results do not suggest rejection of the CCAPM, the estimate of the relative risk aversion parameter greatly exceeds the theoretical value.

[21]The assumptions are (1) joint log-normality of asset returns and consumption growth and (2) a constant relative risk aversion specification of utility.

[22]Since the test assumes both constant relative risk aversion utility and joint log-normality of returns and consumption growth, the violation of either restriction could be the source of the model failure. Further work may be required in this area to decipher the implications of the results.

[23]For a more detailed survey of the APT, see Huberman (1986).

[24]See Sharpe (1985): 199-200.

[25]See Shanken (1982, 1985) and the response by Dybvig and Ross (1985).

[26]The model implications can be extended to the concept of mean-variance efficiency. In the CAPM, the model implies that the market portfolio is mean-variance efficient; that is, given its level of risk, the market portfolio has the highest return. Roll's (1977) critique suggests that this implication has not been tested adequately. For arbitrage pricing theory, a portfolio that has only risk related to the fundamental factors is mean-variance efficient. This aspect of the model has been tested empirically in several studies.

[27]See, for example, Huberman (1982), Chamberlain and Rothschild (1983), and Connor (1984). See Huberman (1986) for a more exhaustive list of references.

[28]Factor analysis is a statistical procedure in which "common factors" are unobservable hypothetical variables that contribute to the variance of a vector of dependent variables. That is, factor analysis is a method to describe the variation of a set of variables without explicit explanatory variables. A data series, then, will be described as a linear function of a set of common factors and one unique factor that contributes variance only to that series. In the set of dependent variables, each variable has one unique factor that is uncorrelated with all other unique factors. The coefficients, or factor loadings, for each common factor provide the estimates of b_i in the APT.

[29]Although factor analysis is more efficient, the computational demands of the method limit the number of returns that can be analyzed at one time.

[30]Huberman (1986) suggests that this form of research, relating expected return to covariances of asset returns with other variables, is more in line with the Merton intertemporal CAPM. Below, the Brock model is used to show the similarity of the two models.

[31]Some references for the earlier yet important works are Roll and Ross (1980), Brown and Weinstein (1983), Chen (1983), and Dhrymes et al. (1985). See Huberman (1986) for a more extensive listing.

[32]Lehmann and Modest (1988) use factor analytic techniques on 750 asset returns to isolate a factor structure. Then they test the APT with these factors on a selection of asset portfolios, grouped on the basis of dividend yield, an asset's own return variance, and firm size. Connor and Korajczyk (1988) estimate factors using asymptotic principal components, which allows more returns in the estimation of the covariance matrix. The asset portfolios used as dependent variables are grouped on the basis of size.

[33]Both papers reject the restriction of mean variance efficiency in APT as well as in the CAPM.

[34]The research that uses prespecified factors to test APT will be examined further below in the discussion of macroeconomic factors and asset pricing.

[35]See Ohlson and Garman (1980) and Connor and Korajczyk (1988) for discussions of intertemporal arbitrage pricing theories.

The intertemporal CAPM, CAPM, and a recent intertemporal model in Cox, Ingersoll, and Ross (1985) have much in common with modern dynamic macroeconomic models (stochastic growth models) such as those of Lucas (1978) and Brock (1982). These similarities include the key role of the real interest rate, concern for the changing investment and consumption opportunities faced by consumers, and the attention to economic forces as the underlying sources of asset risk premia. The progress in modeling a dynamic economy with an asset market has occurred in both fields; empirical work on the issue has numerous potential applications. It is notable, however, that only the ICAPM attempts to isolate underlying economic variables—that is, state variables—directly; the consumption CAPM is similar to the traditional CAPM in the use of a reference portfolio to price assets relative to that portfolio.

[36]McElroy and Burmeister (1988) employ variables similar to Chan, Chen, and Hsieh (1985) and Chen, Roll, and Ross (1986) in a nonlinear estimation method that presents both time-series and cross-sectional pricing results of the multifactor model. Their results generally support the usefulness of the economic variables in explaining both types of variation.

[37]Research by Tallman (forthcoming) investigates the effects of government spending behavior on cross-sectional asset pricing more in the tradition of financial research.

References

Aschauer, David Alan. "Is Public Expenditure Productive?" *Journal of Monetary Economics* 23 (March 1989a): 177-200.

__________. "Public Expenditure and the Stock Market." Federal Reserve Bank of Chicago unpublished manuscript, 1989b.

Bergman, Yacov. "Time Preference and Asset Pricing Models." *Journal of Financial Economics* 14 (March 1985): 145-59.

Black, Fischer. "Capital Market Equilibrium with Restricted Borrowing." *Journal of Business* 45 (July 1972): 444-54.

Black, Fischer, Michael C. Jensen, and Myron S. Scholes. "The Capital Asset Pricing Model: Some Empirical Tests." In *Studies in the Theory of Capital Markets,* edited by Michael C. Jensen. New York: Praeger Publishers, 1972.

Bradley, Michael, and L. Macdonald Wakeman. "The Wealth Effects of Targeted Share Repurchases." *Journal of Financial Economics* 11 (September 1982): 301-28.

Breeden, Douglas T. "An Intertemporal Asset Pricing Model with Stochastic Consumption and Investment Opportunities." *Journal of Financial Economics* 7 (July 1979): 265-96.

Breeden, Douglas T., Michael R. Gibbons, and Robert Litzenberger. "Empirical Tests of the Consumption-Oriented CAPM." *Journal of Finance* 44 (June 1989): 231-62.

Breeden, Douglas T., and Robert Litzenberger. "Prices of State-Contingent Claims Implicit in Option Prices." *Journal of Business* 51 (October 1978): 621-51.

Brock, William A. "Asset Pricing in an Economy with Production: A 'Selective' Survey of Recent Work on Asset-Pricing Models." In *Dynamic Optimization and Mathematical Economics,* edited by Pan-Tai Liu. Vol. 19 of *Mathematical Concepts and Methods in Science and Engineering.* New York and London: Plenum Press, 1980.

__________. "Asset Prices in a Production Economy." In *The Economics of Information and Uncertainty,* edited by J.J. McCall, 1-46. Chicago: University of Chicago Press, 1982.

Brown, Stephen, and Mark I. Weinstein. "A New Approach to Testing Asset Pricing Models: The Bilinear Paradigm." *Journal of Finance* 38 (June 1983): 711-43.

Chamberlain, Gary. "Funds, Factors, and Diversification in Arbitrage Pricing Models." *Econometrica* 51 (September 1983): 1305-23.

Chamberlain, Gary, and Michael Rothschild. "Arbitrage, Factor Structure, and Mean-Variance Analysis on Large Asset Markets." *Econometrica* 51 (September 1983): 1281-1304.

Chan, K.C., Nai-fu Chen, and David A. Hsieh. "An Exploratory Investigation of the Firm Size Effect." *Journal of Financial Economics* 14 (April 1985): 451-71.

Chen, Nai-fu. "Some Empirical Tests of the Theory of Arbitrage Pricing." *Journal of Finance* 38 (December 1983): 1393-1414.

Chen, Nai-fu, Richard Roll, and Stephen A. Ross. "Economic Forces and the Stock Market." *Journal of Business* 59 (July 1986): 383-404.

Connor, Gregory. "A Unified Beta Theory." *Journal of Economic Theory* 34 (October 1984): 13-31.

Connor, Gregory, and Robert A. Korajczyk. "Risk and Return in an Equilibrium APT." *Journal of Financial Economics* 21 (March 1988): 255-89.

Cornell, Bradford. "The Consumption Based Asset Pricing Model: A Note on Potential Tests and Applications." *Journal of Financial Economics* 9 (March 1981): 103-08.

Cowles, Alfred, III. "Can Stock Market Forecasters Forecast?" *Econometrica* 1 (1933): 309-24.

Cox, John C., Jonathan E. Ingersoll, Jr., and Stephen A. Ross. "A Theory of the Term Structure of Interest Rates." *Econometrica* 53 (March 1985): 363-84.

Dhrymes, Phoebus J., Irwin Friend, Mustafa N. Gultekin, and N. Bulent Gultekin. "New Tests of the APT and Their Implications." *Journal of Finance* 40 (July 1985): 659-75.

Dybvig, Philip H., and Stephen A. Ross. "Yes, the APT is Testable." *Journal of Finance* 40 (September 1985): 1173-88.

Fama, Eugene F. "Multiperiod Consumption Investment Decisions." *The American Economic Review* 60 (March 1970): 163-74.

Fama, Eugene F., and James D. Macbeth. "Risk, Return and Equilibrium: Empirical Tests." *Journal of Political Economy* 81 (May/June 1973): 607-36.

Fama, Eugene F., and Merton H. Miller. *The Theory of Finance.* Hinsdale, Ill.: Dryden Press, 1972.

Ferson, Wayne E., and John J. Merrick, Jr. "Non-Stationarity and Stage-of-the-Business-Cycle Effects in Consumption-Based Asset Pricing Relations." *Journal of Financial Economics* 18 (March 1987): 127-46.

Grossman, Sanford J., Angelo Melino, and Robert J. Shiller. "Estimating the Continuous-Time Consumption Based Asset Pricing Model." *Journal of Business and Economic Statistics* 5 (July 1987): 315-27.

Hall, Robert E. "Consumption." National Bureau of Economic Research Working Paper No. 2263 (1987).

Hansen, Lars P., and Kenneth J. Singleton. "Generalized Instrumental Variables Estimation of Nonlinear Rational Expectations Models." *Econometrica* 50 (September 1982): 1269-86.

Hansen, Lars P., and Kenneth J. Singleton. "Stochastic Consumption, Risk Aversion, and the Temporal Behavior of Asset Returns." *Journal of Political Economy* 91 (April 1983): 249-65.

Huberman, Gur. "A Simple Approach to Arbitrage Pricing." *Journal of Economic Theory* 28 (October 1982): 183-91.

__________. "A Review of the Arbitrage Pricing Theory." Center for Research in Security Prices Working Paper Series No. 166 (May 1986).

Hunter, William Curt, and Mary Beth Walker. "Assessing the Fairness of Investment Bankers' Fees." Federal Reserve Bank of Atlanta *Economic Review* 73 (March/April 1988): 2-7.

Jensen, Michael C. "Capital Markets: Theory and Evidence." *The Bell Journal of Economics and Management Science* 3 (Autumn 1972): 357-98.

Keim, Donald B. "Size-Related Anomalies and Stock Return Seasonality: Further Empirical Evidence." *Journal of Financial Economics* 12 (June 1983): 13-32.

Lehmann, Bruce N., and David M. Modest. "The Empirical Foundations of the Arbitrage Pricing Theory." *Journal of Financial Economics* 21 (January 1988): 213-54.

Levy, Haim. "Equilibrium in an Imperfect Market: A Constraint on the Number of Securities in the Portfolio." *The American Economic Review* 68 (September 1978): 643-58.

Lintner, John. "The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets." *The Review of Economics and Statistics* 47 (February 1965): 13-37.

Long, John B., Jr. "Stock Prices, Inflation, and the Term Structure of Interest Rates." *Journal of Financial Economics* 1 (February 1974): 131-70.

Lucas, Robert E., Jr. "Asset Prices in an Exchange Economy." *Econometrica* 46 (November 1978): 1429-45.

Mankiw, Gregory N., and Matthew D. Shapiro. "Risk and Return: Consumption Beta versus Market Beta." *Review of Economics and Statistics* 68 (August 1986): 452-59.

Markowitz, Harry. "Portfolio Selection." *Journal of Finance* 7 (March 1952): 77-91.

McElroy, Marjorie B., and Edwin Burmeister. "Arbitrage Pricing Theory as a Restricted Nonlinear Multivariate Regression Model: ITNLSUR Estimates." *Journal of Business and Economic Statistics* 6 (January 1988): 29-42.

Merton, Robert C. "An Intertemporal Capital Asset Pricing Model." *Econometrica* 41 (September 1973): 867-87.

Merton, Robert C., and Stanley Fischer. "Macroeconomics and Finance: The Role of the Stock Market." In *Essays on Macroeconomic Implications of Financial and Labor Markets and Political Processes,* edited by Karl Brunner and Allan H. Meltzer, 57-108. Carnegie-Rochester Conference Series on Public Policy, vol 21. Amsterdam: North-Holland, 1984.

Mossin, Jan. "Equilibrium in a Capital Asset Market." *Econometrica* 34 (October 1966): 768-83.

Ohlson, James A., and Mark B. Garman. "A Dynamic Equilibrium for the Ross Arbitrage Model." *Journal of Finance* 35 (June 1980): 675-84.

Ritter, Jay R., and Navin Chopra. "Portfolio Rebalancing and the Turn-of-the-Year Effect." *Journal of Finance* 44 (March 1989): 149-66.

Roll, Richard. "A Critique of the Asset Pricing Theory's Tests." *Journal of Financial Economics* 4 (March 1977): 129-76.

Roll, Richard, and Stephen A. Ross. ""An Empirical Investigation of the Arbitrage Pricing Theory." *Journal of Finance* 35 (December 1980): 1073-1103.

Ross, Stephen A. "The Arbitrage Theory of Capital Asset Pricing." *Journal of Economics* 13 (May 1976): 341-60.

Rubenstein, Mark. "The Valuation of Uncertain Income Streams and the Pricing of Options." *Bell Journal of Economics* 7 (Autumn 1976): 407-25.

Schipper, Katherine, and Rex Thompson. "Common Stocks as Hedges against Shifts in the Consumption or Investment Opportunity Set." *Journal of Business* 54 (April 1981): 305-28.

Shanken, Jay. "The Arbitrage Pricing Theory: Is It Testable?" *Journal of Finance* 37 (December 1982): 1129-40.

__________. "Multi-Beta CAPM or Equilibrium-APT?: A Reply." *Journal of Finance* 40 (September 1985): 1189-96.

Shanken, Jay, and Mark I. Weinstein. "Macroeconomic Variables and Asset Pricing: Estimation and Tests." University of Rochester, Simon School of Business, unpublished working paper, 1989.

Sharpe, William F. "Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk." *Journal of Finance* 19 (September 1964): 425-42.

__________. *Investments.* Englewood Cliffs, N.J.: Prentice-Hall, 1985.

Stambaugh, Robert F. "On the Exclusion of Assets from Tests of the Two-Parameter Model." *Journal of Financial Economics* 10 (May 1982): 237-68.

Tallman, Ellis W. "Real Shocks and Asset Premia." Federal Reserve Bank of Atlanta Working Paper (forthcoming).

Tinic, Seha M., and Richard R. West. "Risk Return and Equilibrium: A Revisit." *Journal of Political Economy* 94 (February 1986): 126-47.

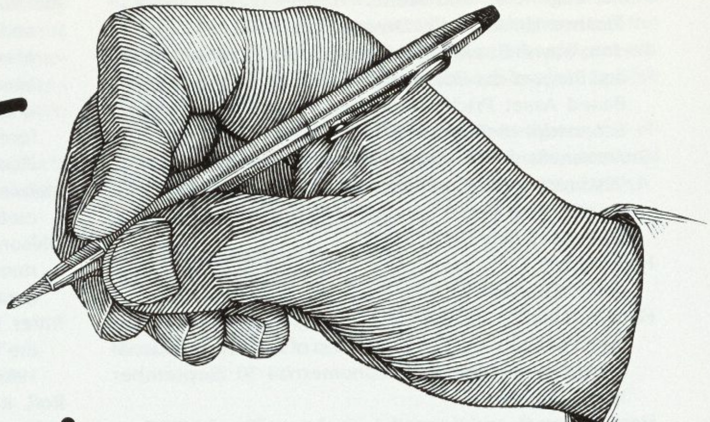
Tobin, James. "Liquidity Preference as Behavior Toward Risk." *Review of Economic Studies* 25 (February 1958): 65-85.

Treynor, Jack. "Toward a Theory of Market Value of Risky Assets." Unpublished manuscript, 1961.

Wheatley, Simon. "Some Tests of the Consumption Based Asset Pricing Model." *Journal of Monetary Economics* 22 (March 1988): 193-215.

Williams, John B. *The Theory of Investment Value.* Cambridge, Mass.: Harvard University Press, 1931.

F.Y.I.



U.S. and Foreign Direct Investment Patterns

William J. Kahley

Since 1985, when this Bank's *Economic Review* last surveyed foreign direct investment, foreign ownership of plants, real estate, and the like in the United States has grown considerably in importance and magnitude. The 1985 article reported: "Many Americans are unaware that . . . international activities wield a sizable and steadily growing impact upon the economic characteristics of the region and the nation." In the meantime both American awareness and direct foreign investment have made significant gains. This article draws on newly available data to update the status of foreign direct investment in the United States and the Southeast. It also probes some economic impacts of foreign direct investment, focusing particularly on employment and exports.

A wide variety of observers, from policymakers to small business owners, are now interested in investments that foreigners make in the U.S. economy. As the magnitude of investment by Europeans, Asians, and others rose in the 1980s, public concern over such foreign investment accelerated, intensifying the demand for information on the subject. Accounts of billion-dollar acquisitions of major U.S. corporations by foreigners often relayed fears that control of corporate America may be slipping out of domestic hands. Opinion polls have shown that the American public is troubled by the increased foreign ownership of U.S. firms and real estate.

In contrast, government officials and other opinion leaders in southeastern states have displayed a generally positive perspective on foreign direct investment. The Southeast receives an especially large share of foreign companies' spending on new plant and equipment. Political and business leaders have come to appreciate the jobs, tax base, and diversification that foreign direct investment can bring.

Many workers also welcome the foreign presence. In 1987 the number of Americans employed in foreign-owned U.S. affiliates was 3.2 million. More than one out of eight of these jobs were in the Southeast, defined in this article as the states in the Sixth Federal Reserve District: Alabama, Florida, Georgia, Louisiana, Mississippi, and Tennessee. Today, southeastern employees working for foreign-owned U.S. affiliates probably number close to half a million.

U.S. ownership of plants abroad generally predates foreigners' direct investments in busi-

---

The author is an economist in the regional section of the Atlanta Fed's Research Department. *He thanks Amy Bailey for extensive research assistance.*

nesses here. The major expansion in U.S. direct investment overseas took place in the 1950s and 1960s, whereas international investments in the United States have been especially fast-growing only since the 1970s. Just as many Americans now oppose foreign direct investment here, critics have faulted U.S. corporations' decisions to produce outside the United States, chiefly claiming that millions of American workers' jobs are lost in the process. For their part, U.S. multinational corporations have argued that their direct foreign presence is needed to serve foreign markets adequately and that customers in the United States benefit from lower prices on imported goods produced at lower-cost foreign subsidiaries.

Based on available data and information, it is impossible to estimate accurately the net economic impact on this country of U.S. companies' foreign investment activities or of foreign direct investment here. However, the ever-expanding amount of research on the topic suggests that both types of flows probably increase U.S. production and well-being. From a theoretical perspective, most economists and policymakers accept the view that international capital flows help companies make better use of the world's resources. This result occurs because foreign investment presumably increases competition in an industry through the entry of new companies. Seeking a competitive edge, firms in an industry try to cut costs, improve efficiency, or enhance product quality to maintain or expand market share. Ultimately, consumers should benefit from a lower-priced or higher-quality product that economizes on resource usage.

This same theoretical perspective also implies that the flow of capital across national borders benefits workers and company owners. If the resources are complementary, employees should be better off because the availability of foreign capital raises labor productivity; consequently, wages should rise or the number of employed workers should increase. The availability of foreign capital also should reduce the cost of capital, making some plant investment projects cheaper and boosting the value of firms, thus benefiting their owners and stimulating investment. On the other hand, domestic savers and financial intermediaries may be losers in the short run as a consequence of greater capital availability. Savers could lose interest income because of lower interest rates brought about by the added supply of capital, while entry of foreign lenders could increase competition in the financial lending business and reduce profitability or the rate of return.

This article presents available information on the magnitude, industrial distribution, and geographic concentration of foreign investment in the United States generally and in the Southeast.[1] Along with a discussion of alternative measures of investment and conceptual and statistical problems associated with these measures, the article compares features of investment in the United States by "U.S. affiliates" of foreign companies with U.S. corporations' investment in "foreign affiliates." The discussion next makes tentative assessments of some impacts of foreign investment in the United States, with particular attention to two questions:

- How, if at all, have job growth and worker income been influenced by foreign investment?
- How, if at all, has foreign investment stimulated U.S. exports?

The final portion of the article focuses on the importance of foreign direct investment in the Southeast. The article concludes with a discussion of emerging trends and prospects for foreign investment in the nation and by U.S. companies during the 1990s.

The Activities of U.S. and Foreign Affiliates

Overall Activity. A statistical snapshot of inward and outward foreign investment for the United States at year's end 1987 reveals some important information. As measured by employment, U.S. multinationals were more active, with 6.2 million workers employed in their foreign affiliates compared to 3.2 million workers in U.S. affiliates of foreign companies (see Table 1).

In contrast, less reliable data on book values of assets suggest that the magnitude of foreign-owned operations in the United States is much closer to that of U.S.-owned operations abroad. The book value of foreign corporations' U.S.

The Meaning and Measurement of Foreign Investment

U.S. statistics-gathering agencies define foreign direct investment in the United States and U.S. investment abroad as ownership or control—directly or indirectly—of 10 percent or more of an enterprise's voting securities, or an equivalent interest by an individual, partnership, group, or organization. Businesses under such control are called *affiliates*, and the investment is said to be *direct*. Although another type of foreign investment in a private enterprise, known as *portfolio investment*, refers to the purchase of stocks or bonds by investors seeking to diversify their assets rather than exercise an effective management role, the terms *foreign direct investment* and *foreign investment* are used interchangeably throughout the rest of this article.[1] In addition, the term *multinational corporation* is used to refer to all foreign investors even though some actually are individuals or other entities.

There are several ways to measure the magnitude and importance of foreign investment. Conceptually, the best measure of importance is annual value added, or contribution to final output. However, value-added data are not available for foreign affiliates, and other measures that only approximate the importance or contribution of foreign investment activity must be used as proxies to compare inward and outward foreign investment.[2] This article focuses on employment and the gross book value of property, plant, and equipment. These complementary gauges lend themselves to calculations of national and regional levels, shares, and growth rates by industry and by country (of origin or destination). Thus, these data serve as proxies for the stock, or cumulative value, of foreign investment and as measures of importance and change in such activity.

Although employment and gross book value data tend to be correlated, or move together, their patterns of change can vary. Therefore, these data series are not equally suited for all purposes. For example, gross book value data do not serve as reliable measures of growth in real industrial activity because they are valued at (constant) acquisition cost. Market value would be better, with values for all years adjusted by prices for some base year. Also, an acquired firm may revalue its assets from historical book value to fair market value, thus changing the asset valuation while employment and the value of production remain static. Thus, data on the number of jobs or employment associated with foreign investment are better for measuring growth in foreign investment activity. On the other hand, the gross book value may be more accurate than employment figures in measuring industrial or regional shares if foreign investment is in capital-intensive industries and industries in which capital has been substituted for labor. Even then, gross book value comparisons are only approximations because of the shortcomings noted. Generally, when the amount of capital used per worker varies from industry to industry, the industrial and regional shares and patterns of change vary. The form that foreign investment takes can also make a difference in terms of its impact. For example, if a merger or acquisition merely involves the purchase of existing assets, the transfer of ownership may generate few or no new jobs. By contrast, capital inflows to build and equip new plants generate new jobs immediately.

Foreign investment activities can be classified according to type and characteristics:

- acquisitions and mergers of enterprises whereby title to stock or assets of a business are secured by a foreign investor;
- a rise in percentage ownership by a foreign investor, known as equity increases;
- joint ventures, in which two or more entities establish a new business according to contractual provisions; and
- new plants and plant expansions, or a foreign investor's establishment of a new operating facility or addition to existing capacity.

Notes

[1] Other major components of foreign investment in the United States include foreign official assets in the United States, such as their holdings of U.S. Treasury securities, and U.S. bank liabilities. Other major U.S. investment assets abroad include U.S. official reserve assets, U.S. government loans, and U.S. bank claims. In 1987 foreign direct investment totaled 26 percent of all U.S. assets abroad and 17 percent of foreign-owned assets in the United States.

[2] Value-added estimates are available for U.S. affiliates at the national level for the 1977-86 period. A comparison of the value added and employment data for these firms for 1986 shows that manufacturing affiliates' employment share of all affiliate employment, 47 percent, was about the same as their share of all affiliates' value added, 45 percent. Value added and employment shares varied widely for other industries.

Table 1.
Selected Data for U.S. and Foreign Affiliates, 1987

	Number of Employees	Millions of Dollars		
		Total Assets	Annual Sales	Annual Employee Compensation
Foreign Affiliates of U.S. Corporations	6,234,600	1,098,166	1,052,260	134,715
U.S. Affiliates of Foreign Corporations	3,159,700	926,042	731,392	93,652

Sources: See U.S. Department of Commerce, Bureau of Economic Analysis (1989a, b, c).

affiliates' property, plant, and equipment was $926 billion in 1987 compared to $1,098 billion for foreign affiliates of U.S. companies. However, U.S. investments abroad are on average much older than foreign-owned investments in the United States and were made when asset prices were far lower. Comparing the magnitude of activity using book values of assets exaggerates the foreign presence in the United States compared to U.S. multinationals' activities abroad. In 1987 both sales and employee compensation of foreign affiliates of U.S. companies substantially exceeded those of U.S. affiliates of foreign companies. These differences also suggest that U.S. outward investment exceeds inward investment.

These employment numbers and income and asset values seem large, but they appear less so when compared to the total national economy. In 1987, employment of foreign multinationals' U.S. affiliates accounted for 3.6 percent of the 86.6 million workers in nonbank U.S. businesses, according to a survey conducted by the U.S. Commerce Department's Bureau of Economic Analysis.[2] This presence doubled the 1.8 percent share recorded in 1977.

Although the overall percentage of workers employed by U.S. affiliates remains small, these shares are significantly larger for some industries. For example, U.S. manufacturing affiliates employed over 7 percent of all U.S. manufacturing workers and accounted for nearly half of the employment by U.S. affiliates in 1987. (Partly because foreign investment is concentrated in chemicals and other industries with relatively low employment-to-assets ratios, U.S. affiliates' 12 percent share of manufacturing assets was larger than the employment share.) Based on available information that shows foreign investment still growing at a fast pace, U.S. affiliates' share in manufacturing employment and assets may well be larger today.

Foreign and U.S. Characteristics. Foreign direct investment here and U.S. firms' direct investments abroad have distinct characteristics. Sources and applications of such investments shown in Tables 2 and 3 display these differences.

The geographic pattern of U.S. direct investment abroad is more dispersed than is the pattern of country sources of foreigners' investments here. Foreign affiliates located in industrialized countries accounted for about 70 percent of employment in U.S.-owned enterprises abroad in 1987, while Canada, Japan, and the European nations accounted for almost 90 percent of employment in all foreign-owned U.S. affiliates.

One of the more noticeable features of foreign investment in the United States over the past decade has been the growing prominence of Japan as a major foreign investor. Between 1977 and 1987 Japan's share of U.S. affiliate employment rose from 6 percent to 9 percent, and its share of assets among U.S. affiliates' rose from 4 percent to 21 percent. Based on 1987 data, Japan ranks first in terms of assets and fourth in terms of employment. However, the relatively recent vintage of Japanese investment may overstate the value of its assets vis-a-vis countries that have long been investing in factories and real estate. Japan also ranks higher by the asset measure than the employment measure because of Japanese investors' acquisi-

Table 2.
Shares of Assets and Employment for U.S. and Foreign Affiliates by Country, 1987

	Employment		Assets	
	U.S. Companies' Foreign Affiliates	Foreign Companies' U.S. Affiliates	U.S. Companies' Foreign Affiliates	Foreign Companies' U.S. Affiliates
Canada	14.6	18.7	13.8	15.2
Europe	41.2	60.2	48.0	50.5
France	5.7	5.8	4.3	3.7
Germany	8.9	11.5	8.3	6.3
Netherlands	2.1	8.5	4.6	7.6
Switzerland	0.8	5.8	3.4	9.0
United Kingdom	12.8	19.9	15.7	16.9
Japan	5.5	9.0	9.7	21.1
Australia, New Zealand, and South Africa	7.2	4.0	4.3	3.0
Latin America	19.7	4.5	14.7	3.5
Middle East	1.6	1.0	2.2	1.9
Other Africa, Asia, and Pacific	9.7	1.4	6.1	2.1

Sources: See U.S. Department of Commerce, Bureau of Economic Analysis (1989b, c).

tions of financial companies, which are asset-intensive.

By industry, manufacturing accounted for 65 percent of foreign affiliates' employment in 1987 but "only" 48 percent of foreign companies' U.S. affiliates' workers were in the factory sector. Retail trade accounted for a much higher share of U.S. affiliate employment, and finance was a much bigger component of U.S. affiliates' assets. Foreigners, particularly from other developed countries, appear strongly attracted both to the large and affluent U.S. consumer market—with its efficient distribution network—and to its financial services industry.

Some other noteworthy contrasts (not shown in the accompanying tables) appear when comparing these investment shares. As might be expected on the basis of differences in country wealth, nations like Italy, Spain, Brazil, Mexico, and many in Asia and the Pacific are more likely to host U.S. investment than to have companies with large interests in the United States. In addition, U.S. foreign investments in developing countries, where wage rates are lower, make considerable use of labor, whereas activity in advanced countries tends to be in more capital-intensive industries.

Comparing Foreign and U.S. Patterns. More detailed analysis of foreign investment data can help determine whether the investment patterns of U.S. and foreign multinational corporations are similar or different. Table 4 shows concentration ratios for manufacturing industries. These ratios are computed by dividing each industry's share of total affiliate employment by the overall U.S. share in the same industry. An industry value greater than one suggests a preference for the industry.

Foreign investment in the United States is concentrated in resource-intensive manufacturing industries, whereas U.S. investment abroad appears to be concentrated in technology-intensive industries. Specifically, U.S. firms' foreign affiliates' concentration ratios exceed one for chemicals; nonelectrical, electronic, and electrical machinery; transportation equipment; instruments; and rubber and plastic products. Of these industries, all but rubber can be classified as technology-intensive. By contrast, in addition to their concentration in chemicals, foreign com-

Table 3.
Shares of Assets and Employment for U.S. and Foreign Affiliates by Industry, 1987

	Employment		Assets	
	U.S. Companies' Foreign Affiliates	Foreign Companies' U.S. Affiliates	U.S. Companies' Foreign Affiliates	Foreign Companies' U.S. Affiliates
Mining	1.6	0.8	1.4	1.3
Petroleum	4.7	3.7	17.9	8.7
Manufacturing	65.4	48.0	38.8	23.6
Food and kindred products	6.5	4.6	3.2	2.5
Chemicals and allied products	9.3	12.2	8.1	8.2
Primary and fabricated metals	4.2	5.0	2.4	2.5
Machinery	19.3	10.2	10.4	3.5
Other	26.1	16.0	14.7	7.0
Wholesale trade	7.9	9.9	9.1	10.5
Retail trade	8.7	18.0	1.4	2.9
Finance, except banking, insurance, and real estate	2.5	6.5	26.0	48.0
Agriculture	1.5	0.4	0.2	0.3
Construction	0.8	1.3	0.4	0.4
Transportation, communication, and utilities	1.4	2.9	1.8	1.0
Services	5.6	8.5	2.9	3.2

Sources: See U.S. Department of Commerce, Bureau of Economic Analysis (1989b, c).

panies' U.S. affiliates have employment concentrations in stone, clay, and glass products; primary metals; and petroleum and coal—industries that are basically resource-intensive.

Impacts of Foreign Investment

One would hope that foreign investment, like domestic investment, promotes economic growth, enhances productivity, and bolsters the competitiveness of U.S. industry. Newspaper accounts have reported anecdotally on some successes of individual foreign investments, such as the revival of moribund U.S. tire companies and the reinvigoration of parts of the automobile manufacturing industry.[3] Besides bringing in new money and possibly increasing net investment and growth, foreign investors also have introduced new technology, such as process engineering, improved quality control methods, and acquainted management with innovative approaches, such as just-in-time inventory systems and quality circles.

Unfortunately, these impacts of foreign investment cannot be quantified systematically. Determining precisely how many U.S. workers' jobs are attributable to foreign investment is not even possible, although direct investment in manufacturing has undoubtedly added jobs in some industries and kept job losses down in others.[4]

Employment impacts are concentrated in manufacturing since U.S. affiliate employment is concentrated there (see Table 2). Among individual manufacturing industries, U.S. affiliates' shares are above total U.S. shares except in textiles, apparel, lumber, furniture and fixtures, rubber, and transportation equipment.

U.S. affiliates appear to be good employers. Compensation per worker at U.S. affiliates increased at an above-average rate in most manufacturing industries compared to compen-

Table 4.
Concentration Ratios for Manufacturing Assets and Sales of U.S. and Foreign Affiliates, 1986*

	Concentration Ratio for Assets		Concentration Ratio for Sales	
	Foreign Companies' U.S Affiliates	U.S. Companies' Foreign Affiliates	Foreign Companies' U.S Affiliates	U.S. Companies' Foreign Affiliates
Chemicals and allied products	2.69	1.95	2.98	1.93
Stone, clay, and glass products	1.89	0.40	2.05	0.34
Primary metal	1.70	0.80	1.92	0.56
Petroleum and coal	1.24	1.01	1.38	1.85
Printing and publishing	0.97	0.10	0.81	0.11
Electrical and electronic equipment	0.94	1.28	1.19	1.17
Food and kindred products	0.79	0.73	0.69	0.69
Paper and allied products	0.67	0.84	0.72	0.74
Fabricated metal	0.63	0.77	0.64	0.60
Instruments and related products	0.53	1.30	0.58	1.25
Machinery, except electrical	0.46	1.51	0.62	1.75
Textile products	0.37	0.44	0.34	0.24
Rubber and plastics products	0.33	1.40	0.37	1.08
Transportation equipment	0.24	1.55	0.35	1.70
Other	0.46	0.89	0.33	0.55

* *Concentration ratios are affiliate export shares of industry sales divided by comparable export shares for all U.S. businesses in an industry.*

Sources: See Howenstine (1988): 59-75, and U.S. Department of Commerce, Bureau of Economic Analysis (1988a, b).

sation increases for all firms in the same industries in the 1977-86 period. Moreover, affiliate compensation per worker was in 1977 already higher in a majority of manufacturing industries, including the important chemicals industry. Compensation per worker for affiliates was also higher in 1986 in every other major employment category (see Table 5) compared to average compensation for all U.S. firms in those industries. However, between 1977 and 1986 compensation per worker grew more slowly in U.S. affiliates than for all U.S. firms in retail trade, agriculture, transportation, communication and utilities, and services, narrowing compensation differentials, though they still favor affiliates.

It is tempting to infer from the foregoing information that foreign investment has improved the compensation of workers at U.S. affiliates compared to all workers in the same industry. However, this conclusion must be qualified. Previous research by the author has shown that foreign direct investment in the United States tends to be attracted to industries populated by large firms; other researchers have established that such firms tend to pay higher wages and offer greater fringe benefits for workers with the same skills, experience, and occupations. When adjusted for these factors, the higher and faster-growing worker compensation observed for U.S. affiliates probably reflects firm and industry size differences in the mix of affiliate versus domestic firms.[5] This finding is consistent with an earlier analysis by the U.S. Commerce Department's Bureau of Economic Analysis. That unpublished study, based on 1980 data covering hourly wages of production workers at U.S. affiliates and all businesses in an industry, concluded that there was no evidence that industrial wage rates for U.S. affiliates and all businesses were different. Also, the U.S. General Accounting Office has compared wages for employees of U.S. affiliates and U.S. automakers and has concluded that wages received were comparable for the two groups.

Proponents of foreign investment often assert that it promotes exports. As a start to

Table 5. Employee Compensation per Worker for U.S. Affiliates and All U.S. Businesses by Industry, 1986

	Annual Employee Compensation per Worker				
	Foreign Companies' U.S. Affiliates		Total U.S.		
	1986	Percent Change 1977-86	1986	Percent Change 1977-86	1986 Ratio of Compensation by U.S. Affiliates to Compensation by all Employers
Mining	$42,849	144.0	$28,899	74.3	148.3
Manufacturing	33,513	106.5	24,441	78.7	137.1
Wholesale trade	32,656	97.0	24,147	75.5	135.2
Retail trade	13,602	37.7	10,988	57.3	123.8
Finance, insurance, and real estate	50,643	227.1	24,857	108.7	203.7
Agriculture	15,949	26.7	13,819	58.7	115.4
Construction	29,855	122.6	22,430	48.4	133.1
Transportation, communication, and utilities	32,213	64.4	26,150	71.9	123.2
Services	18,655	73.0	16,627	81.0	112.2

Sources: See U.S. Department of Commerce, Bureau of the Census (1981, 1988) and U.S. Department of Commerce, Bureau of Economic Analysis (1985, 1988a).

analyzing this assertion, one should answer two questions: are U.S. affiliates concentrated in industries that tend to export, and does foreign investment stimulate exports within a given industry?

The answer to the first question is "probably not." In the American manufacturing sector, nonelectrical machinery, instruments, transportation equipment, and chemicals are the only industries for which exports amount to 10 percent or more of the value of U.S. sales (see Table 6). Of these industries, foreign investment is concentrated only in chemicals, and, as the concentrations depicted in the last column of Table 6 show, U.S. affiliates of foreign chemical companies are less likely to export than are domestic chemical companies.

Industries in which U.S. affiliates had above-average ratios of exports to sales in 1986 relative to all U.S. firms in that same industry included only primary metals, printing and publishing, and petroleum and coal. Overall, U.S. manufacturing affiliates were less likely to export than were U.S.-owned manufacturing firms in 1977 and 1986, and thus the boost to U.S. exports supposedly given by affiliates may not exist. Moreover, the concentration ratios suggest that affiliate manufacturers tended to be less likely to export in 1986 than they were in 1977 compared to their domestic counterparts.[6]

Explanations for these export patterns are not immediately apparent. However, if foreign companies are attracted to producing in the United States primarily to gain access to the U.S. market, affiliate manufacturers might be less likely to export than U.S.-owned manufacturers. Ad hoc explanations for the particular industries' tendencies probably exist also. Unfortunately, lack of detailed information about the ownership and product composition of affiliates in the various industries precludes discussion of such explanations here.

The foregoing analysis suggests that job and worker income growth may have benefited from foreign direct investment. On the other hand, U.S. affiliates' export-generating benefits do not appear positive compared to those for domestic producers.

Table 6.
Exports of U.S. Affiliates and All U.S. Manufacturing Industries
(shares of sales, in percent)

	Foreign Companies' U.S. Affiliates		Total U.S.		Concentration Ratios	
	1977	1986	1977	1986	1977	1986
Manufacturing	5.3	5.7	6.5	7.2	0.83	0.80
Food and kindred products	4.5	2.0	3.5	3.5	1.20	0.58
Chemicals and allied products	6.1	8.8	8.4	10.2	0.72	0.87
Primary metals	5.6	5.5	3.4	4.0	1.68	1.38
Fabricated metals	8.7	4.3	5.7	4.5	1.51	0.95
Machinery except electrical	15.5	9.9	16.0	16.1	0.97	0.62
Electrical and electronic equipment	7.4	4.2	8.0	9.4	0.93	0.98
Paper and allied products	11.9	4.2	4.6	5.4	2.59	0.77
Printing and publishing	2.4	1.5	1.4	1.2	1.69	1.29
Rubber and plastics products	1.9	3.3	4.3	4.9	0.43	0.67
Stone, clay, and glass products	1.7	0.8	3.1	2.9	0.56	0.26
Transportation equipment	9.0	10.0	11.0	11.2	0.81	0.89
Instruments and related products	9.1	7.5	14.0	13.5	0.66	0.56
Petroleum and coal*	4.0	3.7	0.7	1.4	5.82	2.70
Other (textile, tobacco, leather, apparel, lumber and furniture products, and miscellaneous)	12.6	6.1	5.2	5.5	2.41	1.11

* *U.S. affiliates' exports in the petroleum category are assumed to correspond to the petroleum and coal products category in total U.S. exports.*

Sources: See U.S. Department of Commerce, Bureau of the Census (1981, 1988) and U.S. Department of Commerce, Bureau of Economic Analysis (1985, 1988a).

Foreign Investment in the Southeast

Foreign investment's impact in the Southeast also interests analysts and others: Where have the investments been made, and in what industry? What are the impacts of such activity? As with the nation, the task of describing where investment has occurred in the region is fairly simple compared to gauging its economic impacts. Nevertheless, previous analysis by the author of factors motivating foreigners to invest in particular industries and geographic areas has generated some information to help understand the regional impacts of this activity. From these studies of the region, one can conclude that the Southeast has attracted an especially hefty share of such investment and that the region's lures have been favorable business and meteorological climates, above-average economic growth, and plentiful profit opportunities, owing in part to the availability of low-cost resources such as labor and energy.[7]

Comparing aspects of foreign investment in the region with foreign investment nationally using employment data for the 1977-87 period confirms the special favor the Southeast enjoys with foreign investors. Employment growth for all U.S. affiliates during this 10-year span was 159 percent, yet regional affiliate employment growth was 239 percent (see Table 7). This disparity was about as large as the employment growth difference between the region and the nation over the same period. The fast-paced growth of U.S. affiliate employment enabled the region's share of total U.S. affiliate employment to rise from 10 percent in 1977 to more than 13 percent 10 years later.

If regional and national growth rates of U.S. affiliates are compared using book values of assets, a somewhat different and surprising picture is revealed. The nation displays slightly faster growth over the 1977-87 period: 418 percent compared to 378 percent for the region. The disparity between employment and asset-value comparisons may be related to differences in the industry mix of foreign investment at the national and regional levels.

The region's foreign investment differs from the nation's in other ways. To a greater extent than in the United States as a whole, foreign investment in the Southeast has tended to entail building new manufacturing plants and boosting employment-intensive service sector industries such as wholesale and retail trade. Fifteen percent of the value of foreign investment in the Southeast in 1986 was for new plants, compared to less than 6 percent in the nation.[8] Merger and acquisition investments, often involving large capital-intensive U.S. companies, accounted for 61 percent of foreign investment in the region and 63 percent in the nation.

The region continues to be an active host to foreign investment in a wide spectrum of industries and from investors all around the globe. A relatively high share of direct investment in the Southeast has taken the form of new construction in nontraditional industries.

Between 1977 and 1987 affiliate employment growth in the region exceeded comparable growth rates for the nation in all industries except chemicals and real estate. In the chemical industry, regional and national growth rates were equal. In real estate, national affiliate employment growth was more than one-third faster than affiliate growth in the region (see Table 8). Employment shares within the region changed somewhat during this span. The biggest shifts between 1977 and 1986 were a decline in manufacturing's share of affiliate employment to 45 percent from 60 percent in 1977, a rise in retail trade's share from 9 percent to 18 percent, and a rise in the share of other services to 12 percent (from just 0.1 percent). For asset values, industrial affiliate shares within the region shifted similarly.

Reliable data are not available to summarize the impact of employment and asset shifts on regional worker compensation levels or on diversifying the southeastern economy. However, growth of foreign investor interest in the region clearly has shifted toward employment-intensive industries such as services and retail trade. If regional shifts in affiliate versus total

Table 7.
U.S. Affiliate Employment and Growth, 1977-87

	Total Affiliate Employment		
	Amount		Percent Change
	1977	1987	
Alabama	14,313	35,100	145.2
Florida	28,250	116,800	313.5
Georgia	30,693	117,700	283.5
Louisiana	18,367	50,800	176.6
Mississippi	5,734	17,600	206.9
Tennessee	26,215	80,700	207.8
Southeast	123,572	418,700	238.8
United States	1,218,711	3,159,700	159.3

Sources: See U.S. Department of Commerce, Bureau of Economic Analysis (1985, 1989c).

compensation levels by industry followed the national pattern in the 1977-87 period, some relative shift probably occurred out of the generally higher-paying jobs in capital-intensive industries and other manufacturing jobs into lower-paying service sector jobs. However, this shift may also have hastened growth in affiliate employment because the service sector is more labor-intensive.

Differences in national and regional growth rates of employment and assets by industry caused a few significant shifts in the Southeast's industry concentration ratios in the 1977-86 period. Most importantly, the region's concentration in chemicals disappeared, while specialties developed in metals, machinery, and "other" manufacturing industries. In services, an above-average concentration emerged in retail trade, reflecting the especially fast growth of the southeastern market. Shifts in concentration of book values of industry assets showed similar patterns of change.

In 1987 the regional affiliate employment distribution by country of origin of foreign direct investment in the Southeast was similar to the nation's (see Table 9). Latin American and Middle East investors favored the Southeast relative to the rest of the country. Though Japanese investment is often regarded as more widespread in the Southeast than elsewhere, in fact the concentration of Japanese investment is less in the region than in the nation. However,

Table 8.
Foreign Companies' U.S. Affiliate Employment and Growth by Industry, 1977-87

	Southeast		United States	
	Employment Share, 1987 (in percent)	Percentage Change in Employment, 1977-87	Employment Share, 1987 (in percent)	Percentage Change in Employment, 1977-87
Mining	0.6	—	0.9	88.7
Petroleum	4.1	169.5	3.9	35.7
Manufacturing	45.2	153.9	44.3	104.1
Food	3.9	175.5	5.1	124.3
Chemicals	10.3	92.2	12.0	91.8
Metals	5.1	208.5	4.9	81.2
Machinery	11.2	215.4	10.3	103.3
Other	12.9	215.6	12.0	122.2
Wholesale trade	7.9	145.5	9.6	99.1
Retail trade	20.5	665.5	18.3	307.7
Finance	0.6	770.8	1.7	446.0
Insurance	2.6	371.2	2.3	122.3
Real estate	1.6	243.0	1.1	330.8
Other	12.4	—	11.7	351.3
Total	100.0	238.8	100.0	159.3

Components do not add to totals because some detailed data are not published to prevent disclosure of information on individual firms. The effect of data suppression may be to lower the calculated shares for the Southeast in a few instances. Also, meaningful calculation of percentage changes in employment in mining and other industries was not possible.

Sources: See U.S. Department of Commerce, Bureau of Economic Analysis (1985, 1989c).

Japan's southeastern presence has grown rapidly from a substantial base. Japanese affiliates' employment has expanded at about two times the national pace. Assets and sales of Japanese companies' U.S. affiliates have also grown at a substantially faster rate than comparable indicators for other U.S. trading partners. By year's end 1987 more than one job out of ten at Japanese U.S. subsidiaries was in this region.[9]

The major shifts in the Southeast's affiliate shares by country of origin in the 1977-87 period included sharp increases in employment shares for Canada, Japan, Australia, New Zealand, South Africa, and the Middle East, and a drop in share and concentration for Latin America. The latter is undoubtedly related to debt problems in Latin American countries.

Industry and country-of-origin specializations are, of course, linked; for example, European chemical producers own chemical plants across the region. Similarly, the Southeast has been a favorite place for foreign ownership of agricultural land (see Table 10). Much of that ownership is in the region's vast forests, and several of the biggest foreign owners are headquartered in Canada. Currently, Japanese investment interest in Florida citrus cropland and Alabama ranchland is growing, and Japanese activity in the automobile assembly and parts supplying industries is large, particularly in Tennessee.

Emerging Trends and Future Prospects

As shown in this article, foreign investment patterns are exerting a discernible influence on

Table 9.
Foreign Companies' U.S. Affiliate Employment by Country, 1987

	Southeast		United States			
	Number of Employees	Employment Share (in percent)	Number of Employees	Employment Share (in percent)	Southeast Share of U.S.	Southeast Concentration Ratio
Canada	85,500	20.4	590,500	18.7	14.5	1.09
Total Europe	234,100	55.9	1,903,700	60.2	12.3	0.93
France	25,800	6.2	183,600	5.8	14.1	1.06
Germany	30,600	7.3	363,300	11.5	8.4	0.64
Netherlands	39,300	9.4	269,500	8.5	14.6	1.10
Switzerland	17,300	4.1	183,400	5.8	9.4	0.71
United Kingdom	82,800	19.8	630,100	19.9	13.1	0.99
Japan	29,300	7.0	284,600	9.0	10.3	0.78
Other Asia and Pacific	19,500	4.7	149,500	4.7	13.0	0.98
Latin America	34,400	8.2	143,600	4.5	24.0	1.81
Middle East	5,200	1.2	32,500	1.0	16.0	1.21
Africa	1,900	0.5	19,900	0.6	9.5	0.72
United States	3,800	0.9	35,500	1.1	10.7	0.81
All Countries	418,700	100.0	3,159,700	100.0	13.3	—

Components do not add to totals because some detailed data are not published to prevent disclosure on individual firms. The effect of data suppression may be to lower the calculated shares for the Southeast in a few instances.

Source: See U.S. Department of Commerce, Bureau of Economic Analysis (1989c).

the economic landscape of the nation and the Southeast, even if their impact is not easily specified. The current trend toward globalization of markets suggests that multinational corporations will be as prominent during the 1990s as in the decade now concluding. The conditions that have drawn foreign investment to the Southeast do not seem to be disappearing. Foreign investors coming to the United States will probably continue to favor this region. However, several trends are emerging that might dramatically affect future investment flows from abroad to the United States and the Southeast, and vice versa.

The Southeast has been a large and fast-growing segment of the U.S. market. This has helped draw foreign companies as they have looked to locate subsidiary plants. Foreign investors who have built new plants in the region as well as governors of southeastern states (who participate in the groundbreaking ceremonies for many of these plants) have pro-

Table 10.
Land Owned and Mineral Rights Leased or Owned by Foreigners, 1987
(thousands of acres)

	Acres of Land Owned	Acres of Mineral Rights Leased or Owned
Alabama	625	405
Florida	893	737
Georgia	709	70
Louisiana	720	889
Mississippi	369	593
Tennessee	108	98
Southeast	3,424	2,792
United States	13,829	42,531

Source: See U.S. Department of Commerce, Bureau of Economic Analysis (1989c).

claimed the importance of regional growth in attracting investment to the region.

Future Foreign Investment. Several international economic and political developments suggest that inward and outward foreign investment will continue to grow in the 1990s.

- The six-year-old Caribbean Basin Initiative was designed to help Caribbean nations develop their economies by giving them duty-free access to the U.S. market and to encourage American businesses to invest in that region. Now, most U.S. firms can benefit from low labor costs there by operating subsidiaries in the region and exporting to the United States duty-free. Although the apparel industry is excluded from favorable treatment, the very low cost of labor and the requirement to pay duty only on the value added abroad has encouraged growing numbers of U.S. apparel manufacturers to establish operations in the Basin.
- The Mexican government announced in May a liberalization of its direct investment regulations that will permit total foreign ownership of companies with assets of up to $100 million. Mexican officials also promised to remove most restrictions on foreign investment in the tourist industry and to give foreign investors access to previously restricted sectors such as glass, cement, iron, steel, and cellulose.
- The U.S.-Canada Free Trade Agreement, which took effect at the beginning of this year, will phase out all tariffs between the United States and Canada over the next 10 years; ensure "national" treatment so that U.S. and Canadian businesses are free of discriminatory laws at the state, provincial, or municipal level; and loosen Canadian restrictions on U.S. investment. The agreement is likely to boost growth in Canada and spur U.S. investment there.
- The phenomenon that has come to be known as *Europe 1992* will result in a single European Community (EC) market which will replace a dozen separate national markets. The EC now boasts 320 million people with production capacity about equal to that of the United States. If the EC market stays open to foreigners, the standardization brought about by the initiative will enable U.S. companies to operate more freely and efficiently, thereby reducing their production and distribution costs. Removal of geographic barriers will lower transportation costs and encourage development of pan-European marketing efforts, while eliminating technical barriers via uniform regulations and standards should enable companies to reap economies of scale in production.
- Some other countries also have a strong potential for absorbing U.S. foreign investment. South Korea, like Mexico, is liberalizing its investment regulations, although majority Korean ownership will still be required in technologically advanced industries and others deemed critical (such as those which involve large imports of raw materials for processing with a high value added) and defense-related industries. Other developing countries in Asia, Africa, and Latin America offer investment opportunities as well, as do some of the socialist countries, including the Soviet Union under its current policy of openness and reform.

Foreign multinationals' interest in acquiring or establishing American operations also is likely to continue to be strong. Numerous large foreign companies do not yet have a strong direct presence here. Some may want to buy or develop U.S. enterprises that can improve their global market positions. Besides wanting to augment their manufacturing capability, foreign multinationals also are seeking access to new technology and operations that complement existing product lines or furnish a well-known brand name. U.S. and foreign firms also are likely to enter into more partnerships and temporary deals that will increase foreign investment in the United States. All of these factors suggest that the amount of foreign investment is likely to remain high through the next decade.

Foreign investment in the Southeast, in addition to having grown rapidly, has expanded and matured in interesting directions. Geographically, nonmetropolitan areas have increasingly been affected. More and more second-tier companies, especially among Japanese firms, have recently been established to supply larger firms with facilities here or to carve out independent market niches. At the same time growing trade, transportation, and investment linkages between the Southeast and countries around the world are creating new investment opportunities in a broad array of industries.

Even though southeastern states will probably keep benefiting from foreign investment, surveys of affiliate managers have revealed several issues of concern, some of which relate to regional shortcomings. For example, the accounting firm Peat Marwick annually compiles data on foreign-based companies with U.S. headquarters in Georgia.[10] Responses before 1988 identified the lack of quality education in Georgia as one of the top two concerns. (In some instances this problem was perceived by the Peat Marwick authors to be one of image rather than a situation that actually required attention.) Moreover, inadequate labor quality and availability has moved up in importance as a major concern in recent years.

Beyond possible regional drawbacks lie a host of country-specific and some broader international influences that could restrain growth.[11] Protectionist legislation, the tax environment, and the availability of investor incentive programs, for example, are factors of major concern to foreign investors.

Summary

The world economy is in the midst of a direct foreign investment surge. Outlays by foreign investors to acquire or establish U.S. businesses have risen sharply since 1977 and are now producing at record rates. Large foreign multinational corporations are seeking to expand and diversify in world markets, including especially the large, fast-growing, and stable U.S. economy. The Southeast has captured an above-average share of foreign investment, particularly for new plants and activities. Newly developing investment opportunities in the United States and abroad suggest that the globalization process will continue in the 1990s.

Foreign Investment from the Southeast

Just how active internationally are companies headquartered in the Southeast? According to data compiled by the Conference Board, large manufacturing companies based in Alabama, Florida, Georgia, Louisiana, Mississippi, and Tennessee with activities abroad had about $33.4 billion in total sales and $5.4 billion in foreign sales in 1987. Regional companies with foreign investments and with sales greater than $100 million numbered only 32, or about 3.2 percent of all large U.S. multinational corporations. Among the southeastern states, Florida was home to the greatest number of these multinational manufacturers—just 12 firms compared to 150 headquartered in California.

Based on these data, southeastern companies do not appear especially active internationally vis-a-vis the rest of the nation. However, the size distribution of manufacturing firms in the Southeast compared to the rest of the nation is unknown; the region may simply not be the headquarters for many large companies. Moreover, the region could be home to a significant number of companies that are active internationally, but which are small or medium-sized firms.

Although 32 large southeastern-based manufacturing companies hold investments abroad, they do not necessarily have foreign plants. The data show that the region's multinationals own 967 principal U.S. plants yet only 82 foreign plants. Many of the firms may hold licensing agreements with host country companies or simply have sales and service departments abroad.

In general, the southeastern multinationals, which are distributed all over the world, tend to be concentrated in technology-intensive industries. This pattern stands in contrast to foreign investors in the region. These seem to be predominantly European, Canadian, or Japanese and tend to specialize in the resource-intensive industries. A state-by-state summation of activity abroad is presented below.

Alabama. Four companies headquartered in Alabama are active internationally. These four firms have made investments abroad in the clothing and apparel, machinery, electronics, industrial chemicals, concrete, and plastic products industries. Only one owns foreign plants—one in Singapore and one in the United Kingdom.

Florida. All but one of Florida's 12 multinational firms have plants abroad. One manufacturer of general industrial machinery and equipment, as well as optical instruments and lenses, maintains the highest number of foreign facilities, with plants in Belgium, Canada, France, Ireland, and the United Kingdom. Florida's other multinational firms chiefly produce various types of machinery and equipment, electronics, plastics products, and fabricated rubber. The two most common sites for investment seem to be Canada and the United Kingdom.

Georgia. Georgia has the second highest count of multinationals (10), but the largest foreign sales among the region's states. In 1987, multinational corporations headquartered in Georgia earned about $20.3 billion in sales, or three-fifths of total sales by regional multinationals. The Coca-Cola Company, with over half its $7.7 billion sales from foreign markets, has an encompassing global presence. The corporation owns at least 19 foreign plants. Its closest state rival in terms of foreign facilities is a manufacturer of fabrics and carpets with 13 foreign plants in five countries. Popular countries for foreign investment by Georgia's multinationals are Canada, the United Kingdom, and the Dominican Republic.

Louisiana. With only two companies maintaining foreign affiliates, Louisiana hosts the second least number of internationally active firms among the states in the region. Two companies produce construction machinery and equipment and build and repair ships and boats. These two companies have a total of nine foreign plants in Singapore, the United Kingdom, Canada, Egypt, Indonesia, Nigeria, and the United Arab Emirates.

Mississippi. Mississippi appears to be the least active internationally of the southeastern states. In 1987 the state was not credited as home to any large multinational. However, one company, described as the world's largest sound-systems manufacturer, is an example of a smaller-sized firm with international activities. The company operates 17 facilities in the state, as well as a video production studio in Los Angeles, and owns subsidiaries in Canada, England, and the Netherlands.

Tennessee. Tennessee, like Alabama, has a total of four large firms with investments abroad and just two foreign plants among them. The plants are located in Canada and Mexico. The companies are involved in a variety of industries, including machinery and equipment, household furniture, drugs, detergents and cosmetics, and miscellaneous apparel and wood products. Worldwide sales are about $1.1 billion, or just 3.4 percent of total regional multinational sales.

Notes

[1]Discussion of the motivations for foreign direct investment is beyond the scope of this article. Generally, investment motivations are related to expected return, risk, and information considerations. Moreover, certain economic and political forces may help explain longer-run global trends in the types and amounts of foreign investment activity while some other factors influence the precise timing, geographic location among and within countries, and industrial patterns of investment. For detailed discussion of these issues see U.S. Department of Commerce (1984) and Kahley (1987).

[2]The Federal Reserve System regularly collects asset and balance sheet data on the banking industry and, to avoid redundancy, the U.S. Department of Commerce annual surveys do not collect data on banks. However, the Fed does not collect employment data for banks.

[3]Two excellent articles in this vein appeared recently in the *New York Times* (see Hicks, "The Takeover of American Industry" and "Foreign Owners Are Shaking up the Competition," May 28, 1989). These articles describe how foreign companies have altered the competitive dynamics in industries such as chemicals, building materials, tires, automobiles, and steel.

[4]Whether or not there is a net employment gain depends upon the extent to which U.S. affiliates replace U.S. imports (or home country exports) compared to U.S. jobs lost or displaced because of increased competition or because foreign firms' affiliates use more capital at the expense of labor relative to practices of domestic U.S. firms.

[5]There also are some technical problems in comparing affiliate employment and total U.S. employment. At the

detailed industry level, comparisons of employment may not be appropriate because of differences in industry classification between U.S. affiliate and all U.S. business employment data. The affiliate data are classified by industry at the enterprise or company level, whereas total U.S. employment is classified by industry at the establishment level; consequently, affiliate and all industry compensation levels could also be affected by an "industry mix" effect. In addition, U.S. affiliate compensation includes any payments to workers during the year, while employment is as of the end of the year, and the data for total U.S. employment and compensation are as of March.

[6]The statistical finding that U.S. affiliates' likelihood of exporting has dropped while foreign investment has risen sharply in the 1977-86 period may be related to the strong value of the dollar in 1986 compared to 1977. Rather than export from the United States, foreign multinational corporations may have sourced "exports" in some other country, including plants in their home countries.

[7]See Kahley (1985, 1986, 1987).

[8]This information is based on transactions data reported by the U.S. Commerce Department's International Trade Administration.

[9]Data compiled by the Consulate General of Japan in Atlanta show that 496 Japan-affiliated establishments were operating in the Southeast at the beginning of 1989, employing over 41,000 workers. The amount of Japanese investment in the region (excluding Louisiana) at that time was estimated at nearly $5 billion.

[10]See KPMG Peat Marwick (1989).

[11]Generally, the current surge in foreign investment may reflect an attempt by foreign companies to establish a presence in U.S. industries. For example, eight Japanese manufacturers are increasing capacity to be able to produce around 2 million vehicles per year here. To the extent that foreign investment represents an attempt to erase a gap between actual and desired stocks, future investment activity can be expected to slow as the stock adjustment process matures.

References

Burnside, Yvonne. *Key Company Directory.* New York: The Conference Board, August 1988.

Consulate General of Japan. *Japanese Firms, Offices, and Subsidiaries in the U.S. Southeast.* Atlanta, Ga., September 1, 1989.

Consulate General of Japan in New Orleans. *Offices and Subsidiaries of Japanese Firms Within the Jurisdiction of the Consulate General of Japan in New Orleans.* New Orleans, La., April 1, 1989.

Hicks, Jonathan P. "Foreign Owners Are Shaking Up the Competition." *New York Times,* May 28, 1989, section 3, 9.

__________. "The Takeover of American Industry." *New York Times,* May 28, 1989, section 3, 1, 8.

Howenstine, Ned G. "U.S. Affiliates of Foreign Companies: Operations in 1986." *Survey of Current Business.* U.S. Department of Commerce. Bureau of Economic Analysis. Washington, D.C.: U.S. Government Printing Office, May 1988.

Kahley, William J. "Foreign Direct Investment — A Bonus for the Southeast." Federal Reserve Bank of Atlanta *Economic Review* 70 (June/July 1985): 4-17.

__________. "What's Behind Patterns of State Job Growth?" Federal Reserve Bank of Atlanta *Economic Review* 71 (May 1986): 4-18.

__________. "Direct Investment Activity of Foreign Firms." Federal Reserve Bank of Atlanta *Economic Review* 72 (Summer 1987): 36-51.

KPMG Peat Marwick. *Georgia's 1989 Survey of Foreign Companies.* September 1989.

U.S. Department of Commerce. Bureau of the Census. *County Business Patterns, 1978, United States.* CBP-78-1. Washington, D.C.: U.S. Government Printing Office, April 1981.

__________. *County Business Patterns, 1986, United States.* CPB-86-1. Washington, D.C: U.S. Government Printing Office, October 1988.

U.S. Department of Commerce. Bureau of Economic Analysis. *Foreign Direct Investment in the United States: Operations of U.S. Affiliates, 1977-80.* Washington, D.C.: U.S. Government Printing Office, 1985.

__________. *Foreign Direct Investment in the United States: Operations of U.S. Affiliates of Foreign Companies, Preliminary 1986 Estimates.* Washington, D.C.: U.S. Government Printing Office, June 1988a.

__________. *U.S. Direct Investment Abroad: Operations of U.S. Parent Companies and Their Foreign Affiliates, Preliminary 1986 Estimates.* Washington, D.C.: U.S. Government Printing Office, June 1988b.

__________. *Survey of Current Business.* Washington, D.C: U.S. Government Printing Office, July 1988c.

__________. "State Personal Income, 1985-87: Revised Estimates." *Survey of Current Business.* Washington, D.C.: U.S. Government Printing Office, August 1988d.

__________. "U.S. Direct Investment Abroad: Detail for Position and Balance of Payments Flows, 1987." *Survey of Current Business.* Washington, D.C.: U.S. Government Printing Office, August 1988e.

__________. "United States Department of Commerce News." News Release BEA 89-28. Washington, D.C.: U.S. Government Printing Office, June 27, 1989a.

__________. "Selected Data of Foreign Affiliates of U.S. Companies." Computer printouts. July 1989b.

__________. "Selected Data of Nonbank U.S. Affiliates." Photocopies. July 1989c.

U.S. Department of Commerce. International Trade Administration. *International Direct Investment: Global Trends and the U.S. Role.* Washington, D.C.: U.S. Government Printing Office, August 1984.

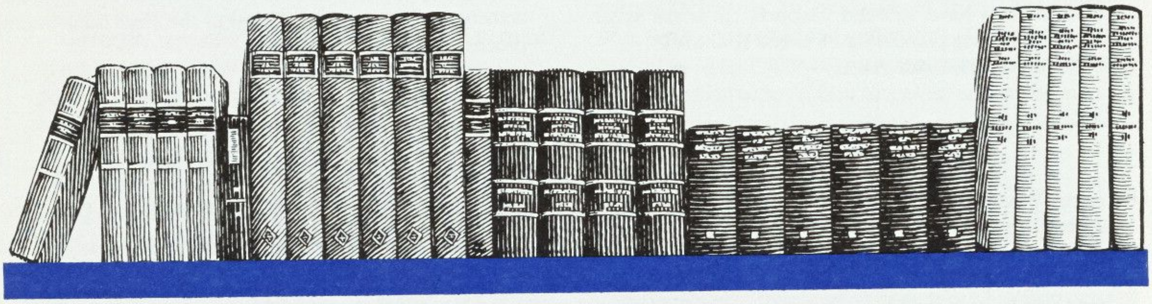
__________. *Foreign Direct Investment in the United States: 1986 Transactions.* Washington, D.C.: U.S. Government Printing Office, September 1987.

U.S. General Accounting Office. *Foreign Investment: Growing Japanese Presence in the U.S. Auto Industry.* Washington, D.C.: U.S. Government Printing Office, March 1988.

Book Review

Bank Costs, Structure, and Performance

by James Kolari and Asghar Zardkoohi.
Lexington, Mass.: D.C. Heath, 1987.
240 pages. $29.00.



Changes in banking regulations in recent years have created new opportunities for commercial banks and other financial institutions to expand their operations. Restrictions on interstate banking and intrastate branching have been liberalized in many states. In addition, legislators and regulators have relaxed many limitations formerly constraining the types of services financial institutions could offer. Along with these developments, though, have come new questions about the future structure of the financial services industry. The industry's composition will depend to a great extent on the types of financial institutions that can remain profitable over time, and profitability will be determined largely by the extent to which banks achieve production economies and resultant cost reductions while expanding their operations.

Two types of production economies are generally available to banks—economies of scale and economies of scope. Economies of scale exist if average production costs decline as output increases. Scope economies are present if two or more products can be jointly produced at a lower cost than is incurred in their independent production. In some industries, such as utilities, it is efficient for a single firm to supply the entire industry output. These industries are termed *natural monopolies* and are characterized by economies of scale at every output level consumers are likely to demand. Recent empirical evidence suggests that the conditions sufficient for natural monopoly in banking are not satisfied over the relevant range of output. Rather, overall economies of scale appear to exist only at low levels of output, and diseconomies, at large output levels, suggesting a U-shaped cost curve for the industry.[1]

Against this backdrop, James Kolari and Asghar Zardkoohi—professors of economics and public policy, respectively, at Texas A&M University—undertake an important and timely task. *Bank Costs, Structure, and Performance* provides a good introduction to the topic of cost economics in banking and an in-depth review of the pre-1985 literature on production economies. The authors reexamine the issue of scope economies in banking and introduce a new measure designed to estimate the extent of these economies. In contrast to past research on bank costs, Kolari and Zardkoohi perform separate analyses (using 1979-83 data) for banks with differing market (product) characteristics. Unfortunately, the researchers fail to emphasize the qualified nature of the evidence they use to draw some of their major conclusions and policy implications. In particular, their cost complementarity and scope economy results should be interpreted with caution. Allen N. Berger addresses these and other shortcomings in another, earlier analysis of Kolari and Zardkoohi's work.[2]

In the first chapter, Kolari and Zardkoohi set the stage for their discussion of costs by describing the external and internal pressures on banking that continue to affect bank profits: increased competition from nonbank institutions, the spread of interstate banking, product line deregulation, and the rapid pace of technological advancement in the industry. These developments have brought about a more competitive environment in which banks must operate as efficiently as possible. The authors' historical overview of twentieth-century banking leads them to suggest that legal and regulatory changes have greatly influenced the structure of U.S. banking. Deregulation has already pushed banks to become more efficient, but its lasting effects cannot be foreseen without some idea of how bank costs behave.

Kolari and Zardkoohi begin their study of bank costs with a review of microeconomic production theory and how it may be applied to the special problems of banks. The authors also delineate sources of scale and scope economies and describe how these are measured. In their discussion of conceptual and methodological issues that arise in estimating such economies, Kolari and Zardkoohi find that the appropriate definition of bank output and choice of functional form are especially important. Accurate measurement of bank output is necessary because economies of scale are defined in terms of the volume of the bank's output.

The researchers analyze the choice of output measure and conclude that dollar values of loans and deposits, rather than number of accounts, should be used. Their argument is that "banks compete to increase market share of dollar amounts as opposed to the number of accounts . . . [and] in a competitive banking environment the cost of an additional dollar of both small and large accounts should be the same." Specification of the functional form of the cost function (and therefore the underlying production function) is also closely related to the measurement of scale and scope effects. The production function identifies the relationship between the quantities of output that result from the use of various quantities of inputs. Comparing three economic production functions—the Cobb-Douglas, constant elasticity of substitution, and translog functions—the authors assert that the last is the preferred form. It is flexible enough to yield U-shaped cost curves (diseconomies as well as economies of scale and scope) and it allows for banks' characteristic multiple inputs and outputs.[3]

Chapter 3 presents a comprehensive survey of previous literature on bank costs, divided into three parts: (1) early studies that relied on simple financial ratios to calculate bank costs, (2) analyses from the mid-1960s and the 1970s that used the Cobb-Douglas function and specified only one output, and (3) more recent works that focus on the translog function. Notwithstanding the various studies' differences in methodologies, output definitions, and data sources, Kolari and Zardkoohi conclude from a review of the earlier studies that "small banks were at a cost disadvantage compared to large banks but that the difference was not so large as to prevent them from competing effectively. . . ." Recent research using the translog cost model yielded results somewhat contrary to those obtained earlier: very small banks were found to be cost-efficient for the most part. All of the studies indicate that most scale economies are exhausted when bank size reaches about $25 million in deposits and that diseconomies of scale exist at large output levels, leading to the familiar U-shaped cost function. However, the evidence on scope economies was ambiguous. Even studies that found positive evidence in favor of joint production concluded that scope benefits were not substantial enough to alter the scale results.

In chapter 4 Kolari and Zardkoohi present the econometric results of their own research. Using Federal Reserve Functional Cost Analysis (FCA) data for 1979-83, they estimate three models: (1) demand deposits and time deposits, (2) loans and securities, and (3) loans and deposits.[4] The dependent variables are the allocated costs for the specific outputs appearing in each regression model. The two researchers find average cost curves to be relatively flat in most cases, so scale is apparently not an important cost determinant. The major implication of a flat cost curve is that many different sizes of banks should be able to coexist. The authors perform jointness tests and find that significant cost complementarities exist only in the joint production of loans and deposits. Kolari and Zardkoohi also use their new measure of scope economies, which gauges the decrease in costs from producing output jointly, as compared to expanding total output by increasing each of the bank's products one at a time (from the minimum level for banks of about the same size). On average, Kolari and Zardkoohi find that banks can reduce expansion costs about 30 percent to 50 percent by increasing outputs at the same time, as opposed to increasing each output separately.

Several issues and problems, both conceptual and methodological, may have influenced the results reported in chapter 4. As a consequence, the usefulness of Kolari and Zardkoohi's conclusions in drawing policy implications, although not eliminated, is limited. First, the FCA data used in the analysis are heavily skewed toward small banks. As of 1986, only 490 banks participated in the program; of this number, 416 held under $200 million in total deposits. To draw conclusions about the cost structure of large banks (over $1 billion in total deposits) based on FCA data is not meaningful and can be misleading. Also, the FCA procedures for allocating costs are sometimes imprecise and may induce bias in the results.

A second problem that may distort the results is that Kolari and Zardkoohi exclude interest payments from their cost measure. Berger, Gerald A. Hanweck, and David B. Humphrey (1987) have shown that studies using dollar measures of outputs and total operating costs as the dependent variable are biased toward finding scale economies because banks can fund a larger asset portfolio by increasing purchased funds. Thus, Kolari and Zardkoohi's analysis is biased by a bank's choice to gather deposits through a branching network or to purchase funds from other retail banks.

A third problem arises with interpreting the authors' cost complementarity and scope economy results. When the results from the translog cost function are used, a *necessary* condition for the existence of cost complementarity between two products is that their cross-product term (δ_{12}) be negative and statistically different from zero. A *sufficient* condition for cost complementarity requires that their cross-product term be not only negative but also greater in absolute value than the product of their output elasticities.[5] However, the cross-product terms reported by Kolari and Zardkoohi are positive in most cases, suggesting that cost complementarity does not hold or, if it does, that negative estimated marginal costs are generating it. Since the authors do not provide the level of complementarities or the estimated marginal costs, it is impossible to determine which condition exists.[6] The fact that Kolari and Zardkoohi do not investigate the scope economy results for statistical significance further detracts from their results.

In chapter 5 the authors test the hypothesis that differences in product mix influence bank cost structures. Based on balance sheet ratios, banks are clustered into four types: farm, retail, city, and wholesale. Only farm banks were found to have unique cost characteristics. (They exhibited flat cost curves where other groups had U-shaped cost curves; they also had scope economies related to deposit size.) All four bank groups had higher scope economies in the joint production of loans and deposits than in the other two models. Kolari and Zardkoohi's results are puzzling, nonetheless. Several researchers who handled differing product mixes by specifying more outputs in the cost function have rejected the hypothesis that different asset and liability categories can be aggregated.[7] Ideally, each bank product should be included as a distinct output, but the availability of data and use of the translog functional form usually limit the level of disaggregation.[8]

Kolari and Zardkoohi turn from static costs to the impact of technological improvements in banking and how they have affected production costs. To test whether larger scale allows more cost-efficient use of technology, the authors regress demand deposit costs on the ratio of computer-related costs to labor costs and some output variables. The closer this ratio is to zero, the greater the cost savings by substituting computer technology for labor. Kolari and Zardkoohi in fact find that the coefficient lies between zero and one, suggesting that cost savings result. They find no evidence of a trend toward greater cost gains by large banks. However, whether this model is sufficiently complete to draw such a conclusion is unclear. In particular, the authors have a very narrow view of technological change and the appropriate costs that are affected. They ignore the possibility that technological innovation can take the form of new production processes rather than equipment. Another problem with Kolari and Zardkoohi's model is that only demand deposit costs are included in the dependent variable.[9]

Finally, the authors examine the relation between cost efficiency and bank failure using Call Report data for 1984. They develop an early-warning-system model based on commonly used financial ratios and individual bank cost measures (scale economies and residual costs) generated in the research reported earlier in the book. The cost measures were found to improve the predictive power of the failure model substantially when added to financial ratios. Problems exist with the analysis, though, because Kolari and Zardkoohi fall into the trap that earlier writers did. By regressing identical operating expenses on loans and deposits separately, the authors' analysis suffers from the same drawbacks as the study by Thomas W. Gilligan and Michael Smirlock (1984): input prices are assumed to be constant and other bank services are excluded even though they affect total operating expenses. This practice gives a bias toward finding both scale and scope economies and leads them to conclude that failing banks were smaller than average.

From a policy perspective, the evidence presented in the book appears to minimize any concern that the banking industry will be dominated by a few large institutions. The lifting of restrictions on interstate banking and intrastate branching might help consolidate resources in states that have limited branch banking and thereby permit small banks to achieve a more efficient scale of production.

On the whole, *Bank Costs, Structure, and Performance* is a useful guide to future work in this area and is of interest to academicians, policymakers, and practitioners. It provides an in-depth look at the literature, introduces a new measure of scope economies, and opens some new lines of research. The book's biggest failing is the absence of necessary qualifications in regard to the econometric evidence it presents and the consequent potential to mislead readers.

Aruna Srinivasan

The reviewer is an economist in the financial section of the Atlanta Fed's Research Department.

Notes

[1]See, for example, Hunter and Timme (1989) and Lawrence and Shay (1986).

[2]See Berger (1988).

[3]The main disadvantage of the Cobb-Douglas production function is that it only allows for uniform scale characteristics, while the constant elasticity of substitution function is highly restrictive in cases where firms produce more than one output or use more than one factor input.

[4]Although FCA data provide information on the number of accounts, Kolari and Zardkoohi prefer to use dollar amounts for the reasons mentioned earlier. In general, researchers take one of two approaches in defining bank costs and output: the production approach or the intermediation approach (Berger, Hanweck, and Humphrey [1987]). Under the production approach, output is measured in terms of the number of loan and deposit accounts, and costs are defined as total operating expenses exclusive of interest costs. The intermediation approach, on the other hand, measures output as the dollar value of the products offered by the bank, and costs include both interest and operating expenses. The intermediation approach uses a broader definition of costs and is considered to be more relevant for addressing issues relating to the long-run viability of banks (Hunter and Timme [1989]).

[5]See Clark (1988).

[6]Other studies (Benston et al. [1983], Mester [1987]) explicitly report negative marginal costs for some products, attributing them to estimation problems such as the presence of multicollinearity and loss of degrees of freedom.

[7]See Kim (1986) and Lawrence and Shay (1986).

[8]See Clark (1988).

[9]It is also important to note that Kolari and Zardkoohi's conclusions are limited to the smaller banks in the economy. Hunter and Timme (1986, 1988) examine the relation between technological change, production economies, and firm size for a sample of large banks and find that larger banks enjoy proportionately higher cost savings from technological change.

References

Benston, George J., Allen N. Berger, Gerald A. Hanweck, and David B. Humphrey. "Economics of Scale and Scope." In *Proceedings of a Conference on Bank Structure and Competition.* Federal Reserve Bank of Chicago (May 1983): 432-55.

Berger, Allen N. In "Book Reviews." *Journal of Money, Credit, and Banking* 20 (May 1988): 283-87.

Berger, Allen N., Gerald A. Hanweck, and David B. Humphrey. "Competitive Viability in Banking: Scale, Scope, and Product Mix Economies." *Journal of Monetary Economics* 20 (December 1987): 501-20.

Clark, Jeffrey A. "Economies of Scale and Scope at Depository Financial Institutions." Federal Reserve Bank of Kansas City *Economic Review* (September/October 1988): 16-33.

Gilligan, Thomas W., and Michael Smirlock. "An Empirical Study of Joint Production and Scale Economies in Commercial Banking." *Journal of Banking and Finance* 8 (March 1984): 67-77.

Hunter, William C., and Stephen G. Timme. "Technical Change, Organizational Form, and Structure of Bank Production." *Journal of Money, Credit, and Banking* 18 (May 1986): 152-66.

Hunter, William C., and Stephen G. Timme. "Technological Change in Large U.S. Commercial Banks." Federal Reserve Bank of Atlanta Working Paper 88-6 (December 1988).

Hunter, William C., and Stephen G. Timme. "Does Multiproduct Production in Large Banks Reduce Costs?" Federal Reserve Bank of Atlanta *Economic Review* 74 (May/June 1989): 2-11.

Kim, Moshe. "Banking Technology and the Existence of a Consistent Output Aggregate." *Journal of Monetary Economics* 18 (September 1986): 181-95.

Lawrence, Colin, and Robert Shay. "Technology and Financial Intermediation in Multiproduct Banking Firms: An Econometric Study of U.S. Banks, 1979-1982." In *Technological Innovation, Regulation and the Monetary Economy,* edited by Lawrence and Shay. Cambridge, Mass.: Ballenger, 1986.

Mester, Loretta J. "A Multiproduct Cost Study of Savings and Loans." *Journal of Finance* 42 (June 1987): 423-45.

Working Paper Series Available

The Research Department of the Federal Reserve Bank of Atlanta publishes a working paper series to stimulate professional discussion and exploration of economic subjects. We welcome readers of the *Economic Review* to complete and return the form below in order to receive our recently released working papers. If you would like a copy of any of the current papers, simply check the box next to the paper's number and return the form to the Public Information Department, Federal Reserve Bank of Atlanta.

89-1 Peter A. Abken
A Survey and Analysis of Index-Linked Certificates of Deposit

89-2 Aruna Srinivasan
Costs of Financial Intermediation under Regulation: Development Banks and Commercial Banks

89-3 William C. Hunter, Stephen G. Timme, and Won Keun Yang
An Examination of Cost Subadditivity and Multiproduct Production in Large U.S. Banks

89-4 Preston J. Miller and William Roberds
How Little We Know about Budget Policy Effects

89-5 Peter A. Zadrozny
Analytic Derivatives of the Matrix Exponential for Estimation of Continuous-Time ARMA Models

89-6 William C. Hunter and Stephen G. Timme
The Demand for Labor at the World's Largest Banking Organizations

89-7 Ellis W. Tallman
Macroeconomic Factors and Asset Excess Returns

89-8 Joseph A. Whitt, Jr.
Nominal Exchange Rates and Unit Roots: A Reconsideration

89-9 Larry D. Wall and David R. Peterson
The Effect of Continental Illinois' Failure on the Financial Performance of Other Banks

Previous working papers are also available. For more information, contact the Public Information Department, Federal Reserve Bank of Atlanta, 104 Marietta Street, NW, Atlanta, Georgia 30303-2713 (404/521-8788).

☐ Please start my subscription to the Working Paper Series.

☐ Please send me a copy of the following working papers:

☐ 89-1 ☐ 89-2 ☐ 89-3 ☐ 89-4 ☐ 89-5 ☐ 89-6 ☐ 89-7 ☐ 89-8 ☐ 89-9

Name ______________________________

Address ______________________________

____

City ______________ State ______________ ZIP ______________

Return to the Public Information Department, Federal Reserve Bank of Atlanta, 104 Marietta Street, NW, Atlanta, GA 30303-2713

BOOK REVIEWS

CORPORATE FINANCE

FEDERAL DEFICIT

FINANCIAL INSTITUTIONS

Index for 1989

FEDERAL RESERVE BANK OF ATLANTA

YEARS 1914-1989

Atlanta Fed History Now Available

In conjunction with the 75th anniversary of the Federal Reserve System, the Atlanta Fed has published a retrospective on its 75 years of service. *A History of the Federal Reserve Bank of Atlanta, 1914 - 1989,* by Richard H. Gamble, recounts the Bank's evolution in response to developments in the region, the nation, and the financial services industry. The story is told through personal recollections, as well as excerpts from official records, and is illustrated with photographs and other items from the Bank's archives.

Beginning with Atlanta's race with New Orleans to become the headquarters for a Reserve Bank, Gamble follows the Atlanta Fed through the Depression, World War II, postwar technological developments, and the Southeast's rapid economic expansion starting in the 1960s. It concludes with a brief recap of the effect of the Monetary Control Act and the challenges of the 1980s.

The illustrated 146-page volume is now available for $4.50 and may be obtained by writing the Public Information Department, Federal Reserve Bank of Atlanta, 104 Marietta Street, N.W., Atlanta, Georgia 30303-2713, or calling 404/521-8268.

Economic Review

Federal Reserve Bank of Atlanta
104 Marietta St., N.W.
Atlanta, Georgia 30303-2713

Address Correction Requested

Bulk Rate
U.S. Postage
PAID
Atlanta, Ga.
Permit 292