Forces That Shape the Yield Curve

MARK FISHER

Fisher is a senior economist and policy adviser in the financial section of the Atlanta Fed's research department. He thanks Lucy Ackert, Christian Gilles, Frank King, Stephen LeRoy, Saikat Nandi, Steve Smith, and Larry Wall for helpful comments on an earlier draft.[1]

MONETARY POLICYMAKERS AND OBSERVERS PAY SPECIAL ATTENTION TO THE SHAPE OF THE YIELD CURVE AS AN INDICATOR OF THE IMPACT OF CURRENT AND FUTURE MONETARY POLICY ON THE ECONOMY. THE YIELD CURVE SHOWS HOW THE YIELD ON A GOVERNMENT BOND DEPENDS ON THE MATURITY OF THE BOND. HOWEVER, DRAWING INFERENCES FROM the yield curve is much like reading tea leaves if one does not have the proper tools for yield-curve analysis. The purpose of this article is to provide a rigorous yet accessible introduction to those tools using high-school algebra.

The yield curve is shaped by expectations of the future path of short-term interest rates and by uncertainty regarding the path. Uncertainty affects the yield curve through two channels: investor attitudes toward risk (risk aversion) as reflected in risk premia and the nonlinear relation between yields and bond prices (known as convexity). In order to present the theory behind the yield curve correctly, uncertainty must be taken seriously. Nevertheless, the source of uncertainty can be modeled quite simply: all uncertainty is resolved by a single flip of a coin. In this setting, all three forces (expectations, risk aversion, and convexity) that shape the yield curve can be rigorously presented. The analysis is organized around the conditions that guarantee the absence of arbitrage opportunities. An arbitrage is a trading strategy that produces something for nothing.

The basic ideas are developed first in an introductory section by the use of an analogy. Next, bond pricing is introduced in a world of perfect certainty, in which no-arbitrage conditions are first worked out algebraically. (In this setting, the absence-of-arbitrage conditions are equivalent to the expectations hypothesis of the term structure of interest rates.[2]) Next, uncertainty is introduced via the coin flip, and the no-arbitrage conditions for bond prices are worked out again. These no-arbitrage conditions are shown to imply the existence of a risk premium that depends on the price of risk (which reflects risk aversion and is the same for all bonds) and the amount of risk (which is measured by the volatility of a bond's price).[3] The last section discusses how to translate the no-arbitrage condition for bond prices into a no-arbitrage condition for yields. The nonlinearity of the price-yield relation brings the convexity term into play.

What Is the Yield Curve?

The simplest kind of bond is called a zero-coupon bond. A zero-coupon bond (also known as a discount bond) makes a single payment on its maturity date. By contrast, a coupon bond makes periodic interest payments, called coupon payments, prior to its maturity when it also makes a final payment that represents repayment of principal. A coupon bond may be thought of as a portfolio of zero-coupon bonds.

A default-free bond is a bond for which all payments are certain to be made in full and on time. U.S. Treasury securities are generally considered to be default-free. The Treasury issues both coupon bonds and zero-coupon bonds. Treasury bills are zero-coupon bonds with original maturities of one year of less. Treasury notes and bonds are coupon bonds with original maturities of two years or more (bonds have original maturities of twenty years or more) that pay interest twice a year. Since the mid-1980s, investors have been able to trade the coupon payments of certain Treasury notes and bonds separately as zero-coupon bonds in what is known as the STRIPS market.[4]

Bonds with different maturities typically have different yields. For example, the yield on a five-year bond is often higher than the yield on a two-year bond. But sometimes the yield on the two-year bond is higher. At any given point in time, the yield curve can be plotted to show the relation between yields and maturity.

Any explanation of the shape of a particular yield curve should be consistent with a combination of expectations, risk premia, and convexity.

In order to focus on the relation between yields and maturity, it is helpful to abstract from a number of other factors that can also affect a bond's yield. For example, bonds issued by private corporations or municipalities (including states and cities) are subject to credit risk, which means simply that the bonds are not default-free. In addition, corporate and municipal bonds are not as actively traded as Treasury securities, and this illiquidity can affect their yields. Some bonds (municipal bonds in particular but also some Treasury securities known as flower bonds) receive special tax treatment.[5] Many bonds, including some Treasury coupon bonds, are callable, which means the issuer has the right to buy them back at a predetermined price at some point in the future. The analysis of bond prices in this article abstracts from all of these factors other than maturity itself. As such, the analysis is most directly applicable to the default-free zero-coupon bonds traded in the STRIPS market.[6]

The Expectations Hypothesis

Historically, the expectations hypothesis has been the most widely used analytical tool to understand the shape of the yield curve.[7] In a nutshell, the expectations hypothesis says that the yield on long-term bonds equals the average of the expected one-period interest rates. If the expectations hypothesis were correct, the slope of the term structure could be used to forecast the future path of the interest rate. For example, if the yield curve were to slope upward at the short end, it would be because the interest rate is expected to rise. One problem with this version of the expectations hypothesis is that in fact the yield curve slopes upward at the short end on average even though interest rates do not rise on average. One way to explain this divergence is to assume that investors are simply wrong on average.[8] But a good theory should not imply that investors are wrong on average.

The expectations hypothesis can be easily modified to account for this persistent upward slope in a way that does not require systematic errors on the part of investors. Since bond prices do fluctuate over time, there is uncertainty (even for default-free bonds) regarding the return from holding a long-term bond over the next period. Moreover, the amount of uncertainty increases with the maturity of the bond. If there were a risk premium associated with that uncertainty, then the yield curve could slope upward on average without implying that interest rates increase on average. If the risk premium were constant, then changes in the slope of the yield curve would forecast changes in the future path of the interest rate. For example, if the slope of the yield curve were to increase, then it would have to be because the path of future interest rates is expected to be higher. This increase in the slope would also imply that future bond yields would be higher. But there is a problem with this version of the hypothesis as well.

Empirical tests of this extended version of the expectations hypothesis (using U.S. data) have shown that changes in the slope of the term structure do a poor job of forecasting changes in the bond yields. In fact, one widely used test shows that an increase in the slope of the yield curve may actually signal a decrease in the future yields. What went wrong in the theory? What went wrong was assuming that the risk premium was constant while in fact it varies over time. Movements in risk premia over time are responsible for a sizable fraction of the movements of the slope of the term structure. When risk premia increase, so does the slope even though expectations are unchanged. As a result, changes in the slope of the yield curve are often negatively correlated with changes in realized yields. It should be noted that the changes in risk premia that bring about this effect can (and do) occur without any change in the risk of the bonds. Risk premia are essentially covariances that change when either the amount of risk or the price of risk changes. In the discussion below, the effects of changing the amount of risk without changing the price of risk will be seen.

There is another feature of the yield curve that the expectations hypothesis has difficulty explaining. The zero-coupon yield curve slopes downward on average at the long end, typically over the range of twenty to thirty years. In other words, the yield on a thirty-year zero-coupon bond is typically below the yield on a twenty-year bond. The expectations hypothesis would suggest that this slope is due to either (1) a persistently incorrect belief that the interest rate will begin to fall about twenty years from now or (2) a decrease in the risk premium for bonds with maturities beyond twenty years, even though the uncertainty of the holding-period return for thirty-year bonds is greater than that for twenty-year bonds. Neither of these reasons is sensible.[9]

There is a sensible explanation, although it may seem counterintuitive at first, for the persistent downward slope of the term structure at the long end. The explanation has to do with the uncertainty regarding the future path of short-term rates. This uncertainty underlies the risk of holding bonds. (If there were no uncertainty regarding the future path, there would be no risk to holding default-free bonds.) Increases in this uncertainty lead (1) to increases in risk premia that increase the slope of the yield curve at the short end and (2) to decreases in the slope of the yield curve at the long end via the effect of convexity. Convexity (technically known as Jensen's inequality) arises from the nonlinear relation between bond yields and bond prices. As a consequence, a symmetric increase in uncertainty about yields raises the average price of bonds, thereby lowering their current yields. This effect is trivial at the short end of the yield curve where it plays no significant role, but it becomes noticeable and even dominant at the long end. The overall shape of the yield curve involves the trade-off between the competing effects of risk premia (which cause longer-term yields to be higher) and convexity (which cause longer term yields to be lower). Typically, the maximum yield occurs in the fifteen- to twenty-five-year maturity range of the zero-coupon yield curve.[10]

This article emphasizes that expectations do in fact play an important role in determining changes in the shape of the yield curve. The reason the expectations hypothesis fails is not that expectations do not matter; rather, the hypothesis fails because it says that nothing else matters. But as has been discussed, the expected future path of interest rates is only one of a number of important forces that shape the yield curve. Any explanation of the shape of a particular yield curve should be consistent with a combination of expectations, risk premia, and convexity.

[1] This article is based in part on a memo written at the Federal Reserve Board coauthored with Christian Gilles.

[2] "Term structure of interest rates" is another way of referring to the yield curve.

[3] This implication—that the absence of arbitrage implies the existence of a risk premium that depends on the price of risk and the amount of risk—is the central message of the article. It is quite general and applies to other asset prices, not just bond prices.

[4] The Treasury STRIPS program was introduced in February 1985. STRIPS is the acronym for separate trading of registered interest and principal of securities. The STRIPS program lets investors hold and trade the individual interest and principal components of eligible Treasury notes and bonds as separate securities.

[5] Taxability is treated in the companion working paper (Fisher 2001).

[6] Even in the STRIPS market, there are other factors at play. Although STRIPS are subject to taxation, once taxes are treated explicitly, the analysis that ignores taxes is essentially correct. Only in the comparison of taxable bonds with tax-exempt bonds is there a need to explicitly account for the effects of taxes. Other factors are more relevant for the internal structure of the STRIPS market. Principal STRIPS often trade at a premium relative to coupon STRIPS because principal STRIPS implicitly contain certain options. Consequently, the analysis presented here is most applicable to coupon STRIPS. (An explanation of the technical reasons for this relationship is beyond the scope of this article.)

[7] Actually there are a number of different but related hypotheses, each of which is called the expectations hypothesis. See Cox, Ingersoll, and Ross (1981) for a discussion of a number of these competing hypotheses. The version described here is the one most often used.

[8] Another way to explain the divergence is to assume that investors give some weight to very large increases in the interest rate that have not yet been observed. This is sometimes called the "peso problem."

[9] There is another explanation—not related to the expectations hypothesis—that is sensible. The downward slope at the long end of the yield curve could, in principle, reflect a substantial demand for the longest-maturity (default-free) zero-coupon bond (for example, to insulate the value of insurance companies' long-term liabilities from interest-rate risk). Although the explanation is not unreasonable, it is unnecessary given the convexity effect discussed below.

[10] It should be stressed that the yield curve typically reported in the newspaper is not the zero-coupon yield curve and may display a somewhat different shape owing to a variety of factors.

No-Arbitrage Conditions: An Introduction

This article has shown that the expectations hypothesis is not a good tool for studying the shape of the yield curve after all, but what will replace it? The fundamental problem with the expectations hypothesis is that it is taken from a world of perfect certainty, in which the expectations hypothesis is a condition for the absence of arbitrage opportunities, and transplanted into a world where there is uncertainty, in which the expectations hypothesis is not a condition for the absence of arbitrage opportunities. Fortunately, in recent years the theory of finance has produced better tools that allow one to directly apply the conditions guaranteeing the absence of arbitrage opportunities in a world where there is uncertainty. The tools were developed as an outgrowth of the famous Black-Scholes model of option prices. The revolution in asset pricing that was initiated by the Black-Scholes model ultimately carried over to bond pricing and the term structure.[11]

It is necessary to have a firm grasp of the no-arbitrage conditions in order to make sense of the shape of the yield curve.

An arbitrage involves trading securities in such a way as to generate something for nothing. Therefore, the conditions that guarantee the absence of arbitrage opportunities have to do with bond prices rather than bond yields. Thus, there is a bit of a paradox: in order to understand the term structure (bond yields), one must move away from the expectations hypothesis (which focuses on yields) and focus instead on bond prices.

The most powerful tool for understanding the term structure of interest rates is called the absence of arbitrage. (This phrase is shorthand for "the conditions that guarantee the absence of arbitrage opportunities.") An opportunity for arbitrage exists when there is an inconsistency in the prices of securities that allows a valuable payoff to be obtained at no cost. For example, if there are two ways to obtain a given payoff and if one way is cheaper than the other, then one can take advantage of this situation by buying the payoff the inexpensive way ("buy low") and selling it the expensive way ("sell high"). The difference is the profit from an arbitrage.[12]

Anyone who prefers more to less would like to take advantage of an arbitrage opportunity. Smart and greedy investors are constantly on the lookout for arbitrage opportunities. In an active and liquid market such as the market for U.S. Treasury securities, any arbitrage opportunities that appear are taken advantage of almost immediately. What happens to an arbitrage opportunity when someone tries to take advantage of it? Buying the payoff in the inexpensive way puts upward pressure on the cost of doing so, and selling the payoff in the expensive way puts downward pressure on the cost of doing so. The result is that an opportunity for arbitrage tends to go away when someone tries to take advantage of it.

In order to understand the conditions that guarantee the absence of arbitrage opportunities, it is useful to think of financial securities as claims to state-dependent payoffs. Different securities contain differing amounts of each possible payoff. Insurance policies are particularly simple in this regard because an insurance policy pays only when a specific state of the world occurs (for example, flood insurance pays only if there is a flood). Other securities may contain a wide variety of payoffs. Derivative securities, such as options, allow for the "disbundling" of the payoffs. For example, one can write a put option on a stock to insure against a fall in its price.

In principle, each of the payoffs in a security's bundle has a separate price. From this perspective, the price of the security is the sum of the (implicit) prices of the payoffs. Here is the key: As long as all of the individual payoffs have positive prices, there will be no opportunities for arbitrage. In other words, arbitrage opportunities arise only if one or more of the payoffs has a zero or negative price. The simplest example of an arbitrage is free insurance. (Free insurance generates something for nothing, but only in some states of the world.) More generally, a trading strategy that generates something for nothing involves buying and selling securities in such a way as to isolate and extract the mispriced payoffs.

These ideas can be illustrated concretely in a mundane setting. Consider a smart shopper at the grocery store. To keeps things simple, suppose the store sells only apples and oranges. Ordinarily when one goes to a store, one sees the posted prices for the produce. If one were to buy a bag containing, for example, two apples and three oranges, the price for the bag of produce would be computed from the prices posted for apples and oranges.

But consider a different kind of store. First of all, apples and oranges are sold mixed together in color-coded grocery bags. There are two combinations available: red bags each contain two apples and three oranges, and blue bags each contain three apples and two oranges. The store posts prices for the bags but not for apples or oranges separately. Even so, a smart shopper can figure out the implicit prices of apples and oranges from the prices of the bags. As long as the implicit prices of apples and oranges are both positive, there will be no arbitrage opportunities. But if the implicit price of either fruit is zero or negative, then one can get something for nothing.

There is another important difference between this store and an ordinary grocery store. Here one can sell bags of produce as well as buy them. For example, if one has two apples and three oranges, one can put them in a red bag (which the store conveniently supplies for free), sell it to the store, and receive the posted price. This repackaging allows a smart shopper who wants only apples to buy only apples. For example, the shopper can buy three red bags (containing a total of nine apples and six oranges), sell two blue bags (containing a total of four apples and six oranges), and end up with five apples left over. The net cost of the apples is the difference between the revenue from selling the two blue bags and the expense of buying the three red bags. Suppose the price of red bags is two dollars and the price of blue bags is three dollars. Then the net cost of apples is zero, and our smart shopper's "trading strategy" involving red and blue bags is an arbitrage: the smart shopper gets something for nothing.[13]

Faced with this arbitrage opportunity, why would the smart shopper limit the size of the trading strategy? Why not buy 3,000 red bags and sell 2,000 blue bags, netting 5,000 apples? Or why not buy three million red bags and sell two million blue bags, netting five million apples? Or why not buy three billion…? The reason, of course, is that at some point the purchases and sales will affect the prices of the bags, driving up the price of a red bag and driving down the price of a blue bag. The changing bag prices will indirectly affect the prices of the apples and oranges, raising the cost of apples. This dynamic reflects the general proposition stated earlier—attempting to take advantage of arbitrage opportunities tends to make them disappear.

How Useful Are No-Arbitrage Conditions?

For some securities, the absence of arbitrage may not be very useful. Consider the prices of Microsoft Corporation stock and Bank of America stock. The absence of arbitrage does not tell us much about the relation between these two stock prices because the state-contingent payoffs that the stocks "contain" do not overlap very much. For a different example, consider the price of Microsoft stock and an option to buy Microsoft stock. In this case, the payoffs are so closely related that the price of the option is completely determined by the no-arbitrage condition (that is, the Black-Scholes model).

The term structure of interest rates is more like the second example than the first. In the stock/option example, there are two risky securities but there is only one source of risk. Similarly for the term structure, there are more bonds than there are sources of risk. Because the payoffs to bonds of different maturities are highly correlated, the absence of arbitrage opportunities is quite useful. On the other hand, as noted above, there is an important difference between the term structure and the stock/option example. In that example, the state of the world is determined by the value of the stock. Because the stock is an asset, the formula for the value of an option is especially simple. In particular, investors' attitudes toward risk play no role. However, for the term structure, the state of the world is determined by the interest rate, and the interest rate is not the value of an asset. Consequently, investors' attitudes toward risk do play a role in the term structure.

Bond Prices and One-Period Returns

The Discount Function. A zero-coupon bond makes a single payment of one unit of payment at some fixed time in the future. For the purpose of exposition, let the unit of payment be the dollar, but the analysis would apply even if the payment were one peso or one "widget." Let $p(t, n)$ be the value at time t of a zero-coupon bond that matures at time $t + n$, where n is the term to maturity of the bond. Holding t fixed and varying n in $p(t, n)$ traces out the discount function at time t. The value of a zero-coupon bond tells how much a risk-free payment paid in the future is worth today. Two properties of bond prices

[11] See Black and Scholes (1973). In the Black-Scholes model, the stock price summarizes the "state of the world" for option prices. In modeling the term structure, it is the interest rate, rather than the price of an asset, that summarizes the state of the world for bond prices. This difference accounts for the time lag in adapting the Black-Scholes paradigm to bond prices.

[12] This example highlights the fact that when the "law of one price" is violated, an arbitrage opportunity exists.

[13] In order to avoid arbitrage opportunities, the ratio of the cost of blue bags to the cost of red bags must be greater than two-thirds and less than three-halves. In this example the ratio was exactly three-halves, which allows arbitrage opportunities.

are immediately apparent. First, the value of one dollar to be delivered immediately is one dollar; that is, $p(t, 0) = 1$ (see the table). Second, the value of a dollar to be delivered in the infinite future is zero; that is, $\lim_{n \to \infty} p(t, n) = 0$.[14] Chart 1 shows a discount function.

One-Period Returns. Suppose one buys an n-period bond today and sells it next period when it becomes an $(n - 1)$-period bond. The bond that costs $p(t, n)$ today can be sold for $p(t + 1, n - 1)$ next period, as shown in the table. The holding-period return for this investment is

$$\frac{p(t+1, n-1)}{p(t, n)} - 1 = \frac{p(t+1, n-1) - p(t, n)}{p(t, n)},$$

which is the amount one has at the end of the period divided by the amount one invested at the beginning of the period minus one.

In general, it is not known in advance what the price of an $(n - 1)$-period bond will be in the next period, and consequently the holding period return is uncertain. The central point of this article is to uncover the relation between the average holding period return and this uncertainty.

For now, focus on the holding-period return on a one-period bond, which *is* known in advance since the one-period bond delivers one dollar without fail next period (see the table). This return can be defined as the one-period risk-free interest rate. A one-period bond can be purchased today for $p(t, 1)$. The amount repaid next period equals the amount loaned plus interest:

$$1 = [1 + r(t)]\, p(t,1). \qquad (1)$$

Equation (1) can be solved for the one-period risk-free interest rate:

$$r(t) = \frac{1}{p(t,1)} - 1.$$

Today's Price: The Present Value of Next Period's Price

The relation between bond prices today and bond prices in the next period is examined below. This examination involves forming a portfolio today that costs nothing and finding out what it will be worth in the next period. An n-period bond will be purchased and financed by borrowing its cost at the one-period risk-free rate. (In other words, one-period bonds of equal value will be sold.) The net cash flow at time t is zero. Next period, the long-term bond will be sold and the debt repaid (principal plus interest). The table summarizes the net cash flows for this trading strategy.

Net Cash Flows

Buying an *n*-Period Bond and Holding until Maturity

Today (t)	At maturity ($t + n$)
$-p(t, n)$	1

Buying an *n*-Period Bond and Holding One Period

Today (t)	Next period ($t + 1$)
$-p(t, n)$	$p(t + 1, n - 1)$

Buying a One-Period Bond

Today (t)	Next period ($t + 1$)
$-p(t,1)$	1

Financing the Purchase of an *n*-Period Bond with One-Period Borrowing

Today (t)	Next period ($t + 1$)
0	$p(t + 1, n - 1) -$ $[1 + r(t)]\, p(t, n)$

If it is known today that $p(t + 1, n - 1)$ will be greater than $[1 + r(t)]p(t, n)$, then the trading strategy is an arbitrage: something (next period) is obtained for nothing (today). On the other hand, if it is known today that $p(t + 1, n - 1)$ will be less than $[1 + r(t)]p(t, n)$, the trading strategy can be modified to make it an arbitrage. Instead of buying the long-term bond and selling some one-period bonds, sell the long-term bond and buy the one-period bonds. The net cash flows for this trading strategy are the same as for the original trading strategy except that the signs are reversed. The upshot is that in a world of no uncertainty, the absence-of-arbitrage condition for bond prices is

$$p(t + 1, n - 1) - [1 + r(t)]\, p(t, n) = 0. \qquad (2)$$

Equation (2) can be solved for today's price of the long-term bond:

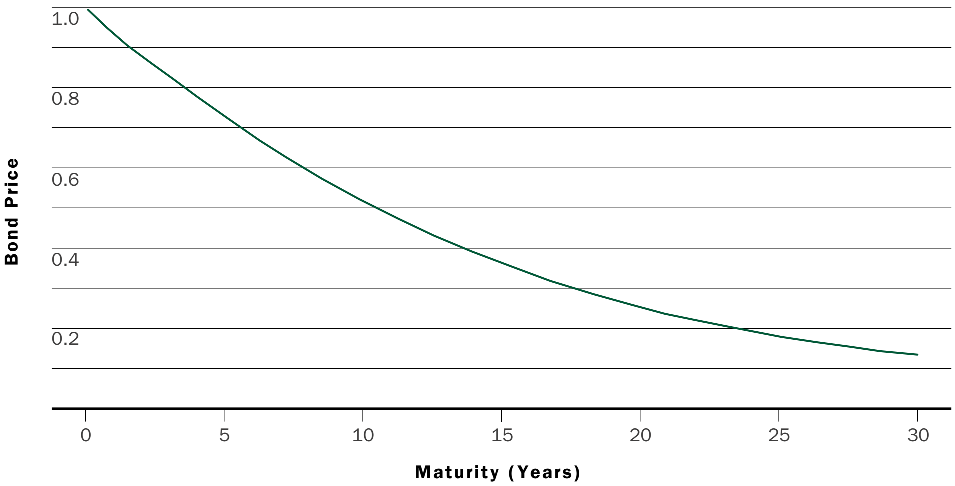
$$p(t, n) = \frac{p(t+1, n-1)}{1 + r(t)}. \qquad (3)$$

In other words, the price of the bond today is the present value of its price in the next period. Another way to express this is

$$\frac{p(t+1, n-1) - p(t, n)}{p(t, n)} = r(t),$$

which says that the (net) return on a bond equals the risk-free interest rate.

CHART 1 The Discount Function: The Price of Zero-Coupon Bonds



Uncertainty

All of the bonds dealt with in this article are default-free—that is, all promised payments are made in full and on time. Nevertheless, these bonds have risk prior to maturity: they can gain or lose value. This uncertainty regarding bond prices can (and will) be linked to the uncertainty regarding interest rates, and this latter uncertainty can be viewed as more fundamental. Nevertheless, the effect of that uncertainty on bond prices and on the conditions that guarantee the absence of arbitrage opportunities can be studied without reference to the underlying interest rate uncertainty.

In the previous section, an absence-of-arbitrage condition based on knowing next period's bond value with certainty was established. (See equation [2].) What if the bond's value in the next period is not known with certainty? What if its possible values can make the net cash flow for a trading strategy sometimes positive and sometimes negative? In this case, the trading strategy is not an arbitrage. The conditions for the absence of arbitrage opportunities are not sufficiently restrictive to completely establish the relation between today's price and next period's price when there is uncertainty. Nevertheless, they do put enough structure on bond prices to provide useful results.

Heads or Tails? All bond prices tend to go up and down together. When the short-term interest rate rises, all bond prices tend to fall, and, conversely, when the short-term interest rate falls, all bond prices tend to rise. To keep things simple, suppose there are only two possible discount functions in the next period. The flip of an unbiased coin will determine which discount function is realized.[15] In other words, if one were to buy an n-period bond today, there would be two possibilities for the price of an $(n - 1)$-period bond in the next period, with the actual outcome determined by the flip of a coin. The notation can be simplified a bit by limiting consideration to just today (time t) and tomorrow (time $t + 1$). Let the price today of an n-period bond be p_n. (See the appendix for a list of the notations used in this article.) If the coin comes up heads, the price of the bond tomorrow will be p^H_{n-1} and if it comes up tails the price will be p^T_{n-1}. Let $\bar{p}_{n-1}$ denote the average price of the bond in the next period:

$$\bar{p}_{n-1} = \frac{p^H_{n-1} + p^T_{n-1}}{2}.$$

Let δ^p_{n-1} denote the volatility of the bond price in the next period:

$$\delta^p_{n-1} = \frac{p^H_{n-1} - p^T_{n-1}}{2}.$$

Volatility is a measure of the riskiness of the investment. It is related to the variance and the standard

[14] This property holds if the interest rate is always positive. If the interest rate can be negative, then the discount function does not have to go to zero. So-called nominal interest rates cannot take on negative values because one can always hold currency instead (which has a nominal return of zero).

[15] An unbiased coin has a fifty-fifty chance of coming up either heads or tails.

deviation. The variance is the average squared deviation from the mean,

$$\frac{1}{2}(p^H_{n-1} - \overline{p})^2 + \frac{1}{2}(p^T_{n-1} - \overline{p})^2 = (\delta^p_{n-1})^2,$$

and the standard deviation is the square root of the variance, which is the absolute value of the volatility, δ^p_{n-1}. Volatility is more useful than standard deviation because volatility's sign plays a role in characterizing whether the risk is bad or good. (An insurance policy is an example of an investment that has good risk because it pays off in bad times). The two possible values of the $(n - 1)$-period bond in the next period as determined by the coin flip are $p^H_{n-1} = \overline{p}_{n-1} + \delta^p_{n-1}$ and $p^T_{n-1} = \overline{p}_{n-1} - \delta^p_{n-1}$. Chart 2 plots two postflip discount functions and their average.

Although there is no need to specify which of the two postflip prices is greater, for the sake of concreteness assume $p^H_{n-1} > p^T_{n-1}$ and therefore $\delta^p_{n-1} > 0$.

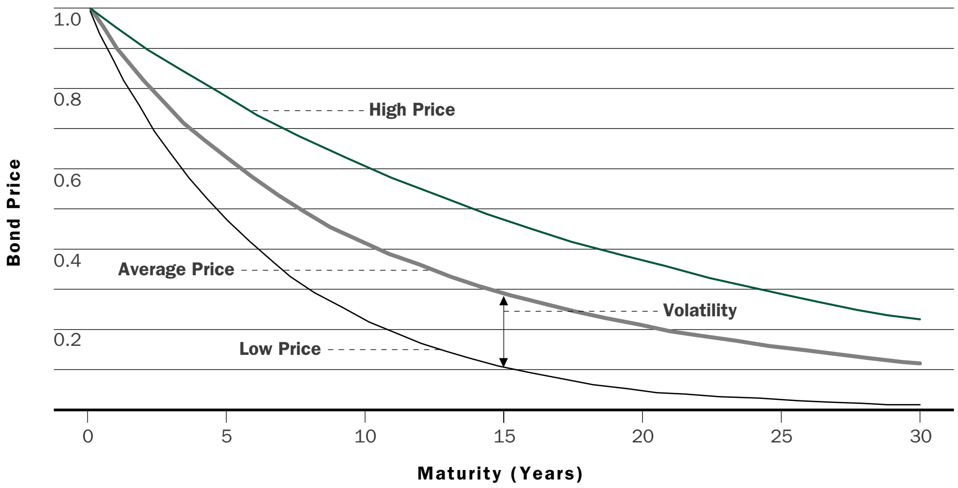
The Absence of Arbitrage Opportunities under Uncertainty: Part I. Recall that an arbitrage is a trading strategy that generates "something for nothing." Now that uncertainty has been introduced, what the absence of arbitrage means needs to be reexamined. Suppose there were a trading strategy that had zero net cash flow today. In other words, the trading strategy costs nothing. The conditions for absence-of-arbitrage opportunities can be stated in terms of the net cash flows next period as follows: Either (1) they are both zero (as they must be in the case of no uncertainty) or (2) one is positive and the other is negative.

To see why these conditions must be so, suppose the contrary were true. If, for example, the net cash flows next period were both positive, then the trading strategy would clearly generate an arbitrage: one would get something in the next period—in all states of the world—for nothing today. On the other hand, suppose only one net cash flow were positive next period and the other were zero. This situation too would generate an arbitrage: just like free insurance, it would cost nothing today and make positive payoffs in some states of the world next period, without the possibility of negative payoffs. Alternatively, if both payoffs were negative (or one negative and the other zero), one could reverse the signs of the payoffs by reversing the positions in the trading strategy (for example, selling instead of buying, lending instead of borrowing).

This analysis can be applied to the following simple trading strategy: Buy an n-period bond today and finance its purchase price with one-period risk-free borrowing. The net cash flow today is zero, and the two possible net cash flows in the next period are $p^H_{n-1} - (1 + r)p_n$ and $p^T_{n-1} - (1 + r)p_n$. If there were no uncertainty ($p^H_{n-1} = p^T_{n-1}$), the no-arbitrage condition would be that both net cash flows in the next period must be zero. But when there is uncertainty ($p^H_{n-1} \neq p^T_{n-1}$), the two net cash flows cannot both be zero. In this case, the no-arbitrage condition is that $(1 + r)p_n$ must lie between p^H_{n-1} and p^T_{n-1}, thereby guaranteeing that one net cash flow is positive and the other negative.[16]

Today's Price: The Present Value of Next Period's Adjusted Average Price. By aping the

CHART 2 Two Postflip Discount Functions and Their Average



Note: The volatility is a measure of the uncertainty.

relation between today's price and next period's price that was established when there was no uncertainty, some guidance in how to proceed can be obtained. The simplest and most natural way to modify equation (3) so that it makes sense when the value of a bond next period is not certain is to replace the uncertain price next period with its average:

$$p_n = \frac{\overline{p}_{n-1}}{1+r}, \tag{4}$$

where $r = r(t)$. Equation (4) says that today's bond price is the present value of the "expected value" of tomorrow's bond price.[17] Equation (4) can be written as

$$\frac{\overline{p}_{n-1} - p_n}{p_n} = r,$$

which says that the expected return on a long-term bond equals the risk-free rate (that is, the risk-free return on a one-period bond).

But why should investors be willing to earn exactly the risk-free rate on average? If the uncertainty associated with owning bonds contributes to the overall uncertainty of investors' lives, investors may require a higher average return to take on this additional risk. On the other hand, if the uncertainty associated with owning bonds reduces the overall uncertainty of their lives, they may accept an average return that is less than the risk-free rate.

In order to account for how investors feel about the kind of risk they face, an adjustment term (a_{n-1}) can be incorporated into the formula for today's bond price:

$$p_n = \frac{\overline{p}_{n-1} - a_{n-1}}{1+r}. \tag{5}$$

The numerator on the right-hand side of equation (5), $\overline{p}_{n-1} - a_{n-1}$, is referred to as the adjusted average price. Equation (5) says that today's price is the present value of next period's adjusted average price. Equation (5) can be rearranged to express the expected return for the bond:

$$\frac{\overline{p}_{n-1} - p_n}{p_n} = r + \frac{a_{n-1}}{p_n}. \tag{6}$$

Equation (6) says that the average holding-period return for a bond is the risk-free rate plus an additional term that somehow accounts for the amount and type of risk involved.

The adjustment term, which can be positive, negative, or zero, provides great flexibility within certain bounds. It has already been shown that $(1 + r)p_n$ must be between p^H_{n-1} and p^T_{n-1} in order to avoid arbitrage opportunities. Given equation (5), these boundaries imply that the adjusted average price, $p_{n-1} - a_{n-1}$ must also be between p^H_{n-1} and p^T_{n-1} (since $[1 + r]p_n$ equals the adjusted average price). Within these bounds, any bond price (or expected return) can be obtained with a suitable choice for the adjustment term. Stated differently, no bond prices in this range can be ruled out. In other words, thus far the theory of bond pricing under uncertainty provides very little structure. To obtain more structure, the way in which two different long-term bonds interact must be examined.

The Absence of Arbitrage Opportunities under Uncertainty: Part II. This section examines arbitrage opportunities that involve simultaneously buying and selling bonds with different maturities in order to form a risk-free portfolio.[18] In the course of this examination, the condition that guarantees the absence of arbitrage opportunities will be uncovered. As will be seen, that condition has something important to say about how the adjustment terms on different bonds relate to each other.

Consider the following portfolio of two bonds: Buy one n-period bond and buy (or sell) some m-period bonds (where m is different from n). Let b denote the number of m-period bonds purchased (where b is negative if they are sold). The cost of this portfolio today is $p_n + bp_m$, which may be positive, negative, or zero. Let π^H and π^T represent the possible values of this portfolio next period. The two values are $\pi^H = p^H_{n-1} + bp^H_{m-1}$ and $\pi^T = p^T_{n-1} + bp^T_{m-1}$.

Each of these two bonds is risky in isolation. But since the uncertainty for each of these bonds is driven by the same underlying source of risk, it is possible to combine the bonds in such a way as to reduce the overall risk. In fact, there is a value for b (call it b^*) that makes the portfolio completely risk-free. In other words, the value of the portfolio next period is the same in both states of the world

[16] This condition guarantees that the realized return on the n-period bond is greater than r if the coin comes up heads and less than r if it comes up tails.

[17] Equation (4) is an expectations hypothesis, albeit one based on bond prices rather than on interest rates. In the companion working paper (Fisher 2001), the typical statement of the expectations hypothesis is discussed, namely, that forward rates are expectations of future one-period returns.

[18] See Vasicek (1977) for an early application of the absence of arbitrage to the term structure of interest rates.

so that $\pi^H = \pi^T$. For this statement of value to be true, b^* must satisfy

$$p^H_{n-1} + b^* p^H_{m-1} = p^T_{n-1} + b^* p^T_{m-1}. \qquad (7)$$

Equation (7) can be solved for

$$b^* = -\left(\frac{p^H_{n-1} - p^T_{n-1}}{p^H_{m-1} - p^T_{m-1}}\right) = -\left(\frac{\delta^p_{n-1}}{\delta^p_{m-1}}\right). \qquad (8)$$

Since b^* is negative, this portfolio involves selling some m-period bonds. In other words, b^* is a hedge ratio—it tells us how to use one bond to hedge the risk of another so that on balance there is no risk at all.[19] Let π^* denote the known payoff to this risk-free portfolio. Since π^* can be computed from either side of Equation (7), it must equal the average of the two sides:

$$\pi^* = \overline{p}_{n-1} + b^* \overline{p}_{m-1}.$$

The cost of the risk-free portfolio is $p_n + b^* p_m$. Consider a trading strategy in which the risk-free portfolio is financed with one-period borrowing. Since the net cash flow today is zero and the net cash flow next period is certain, there will be an arbitrage opportunity unless the cash flow next period is zero. Therefore, the condition for the absence of arbitrage opportunities is

$$\pi^* - (1 + r)(p_n + b^* p_m) = 0. \qquad (9)$$

In order to see what this condition implies for the adjustment terms of the two bonds, equation (5) can be used to re-express the cost of this portfolio using the adjusted average prices:

$$\begin{aligned} p_n + b^* p_m &= \overbrace{\left(\frac{\overline{p}_{n-1} - a_{n-1}}{1+r}\right)}^{p_n} + b^* \overbrace{\left(\frac{\overline{p}_{m-1} - a_{m-1}}{1+r}\right)}^{p_m} \\ &= \frac{\overbrace{\left(\overline{p}_{n-1} + b^* \overline{p}_{m-1}\right)}^{\pi^*}}{1+r} - \frac{\left(a_{n-1} + b^* a_{m-1}\right)}{1+r} \\ &= \frac{\pi^*}{1+r} - \frac{\left(a_{n-1} + b^* a_{m-1}\right)}{1+r}. \qquad (10) \end{aligned}$$

At this point $p_n + b^* p_m$ can be replaced in the no-arbitrage condition (9) by the last line on the right-hand side of equation (10), so that the no-arbitrage condition becomes

$$a_{n-1} + b^* a_{m-1} = 0. \qquad (11)$$

Equation (11) shows that the adjustment terms play a central role in the condition that guarantees the absence of arbitrage opportunities.

The final expression for the absence-of-arbitrage condition can now be found. Substituting the solution for b^* given in equation (8) into equation (11) and rearranging produces

$$\frac{a_{n-1}}{\delta^p_{n-1}} = \frac{a_{m-1}}{\delta^p_{m-1}}. \qquad (12)$$

Equation (12) says that the ratio of the adjustment term to the bond-price volatility must be the same for both bonds. This common ratio is called the price of risk. Let λ denote the price of risk, so that

$$\lambda = \frac{a_{n-1}}{\delta^p_{n-1}} = \frac{a_{m-1}}{\delta^p_{m-1}}.$$

The absence-of-arbitrage condition does not say whether the price of risk is big or small or even whether it is positive, negative, or zero; it says only that it must be the same for all bonds.

The Adjustment Term Is the Risk Premium. Given the absence-of-arbitrage condition just established, the adjustment term can be written as

$$a_{n-1} = \lambda \delta^p_{n-1}, \qquad (13)$$

where λ is the price of risk and δ^p_{n-1} is the volatility of the bond's price. Equation (13) can be expressed as *risk premium = price of risk × amount of risk*. In other words, the adjustment term is the risk premium and the volatility of the bond price is the amount of risk that earns a premium.

The condition for the absence of arbitrage opportunities can be stated in terms of the expected return on a bond by substituting equation (13) into equation (6):

$$\frac{\overline{p}_{n-1} - p_n}{p_n} = r + \lambda\left(\frac{\delta^p_{n-1}}{p_n}\right), \qquad (14)$$

where δ^p_{n-1}/p_n is the relative volatility of the bond price; it measures the volatility of the holding-period return. Equation (14) can be expressed as *expected return = risk-free rate + (relative) risk premium*, where the relative risk premium equals the price of risk times the amount of risk as measured by the relative volatility of the bond price. In other words, the extra return one gets (from the risk premium) depends on the amount of risk (δ^p_{n-1}/p_n) and the price of risk (λ). If either is zero, there is no risk premium.[20]

Bond Yields and Convexity

In this section, the yield to maturity is defined and the absence-of-arbitrage conditions are reexpressed in terms of yields.

Yield to Maturity. Suppose an n-period bond were purchased. If it were held until it matured, what would the return on the investment be? If the amount invested were $p(t, n)$ and the amount returned were one, the total gross return would be simply

$$\frac{1}{p(t,n)}.$$

From the total gross return, the gross return per period could be computed:

$$p(t,n)^{-1/n} = \frac{1}{\sqrt[n]{p(t,n)}}$$

since

$$\overbrace{\frac{1}{\sqrt[n]{p(t,n)}} \times \frac{1}{\sqrt[n]{p(t,n)}} \times \cdots \times \frac{1}{\sqrt[n]{p(t,n)}}}^{n \text{ times}} = \left(\frac{1}{\sqrt[n]{p(t,n)}}\right)^n = \frac{1}{p(t,n)}.$$

Typically, however, it is not the gross return period that is used to characterize the return but rather the net return per period. The net per-period return is called the yield to maturity (or simply the yield). The yield is like an "interest rate." There is a degree of freedom in computing interest rates: how many times per period is interest assumed to be compounded? The fact that there are only two points in time under consideration (the beginning of the period and the end of the period) does not resolve the issue since one is free to quote the interest rate as if there were subperiods over which compounding takes place. Let $y^i(t, n)$ denote the value of y that solves the following equation for a given i: $(1 + y/i)^i = p(t, n)^{-1/n}$ $i = 1, 2, 3, \cdots$. The solution is $y^i(t, n) = i[p(t, n)^{-1/(ni)} - 1]$. Given the price of the bond, each and every $y^i(t, n)$ has a right to be called the net return per period. How one chooses to quote the return (that is, the value one chooses for i) is merely a matter of convenience.

There are two rates of compounding that are particularly convenient to use, and they happen to lie at opposite ends of the compounding spectrum. The first case is called simple compounding, where interest is compounded only once per period ($i = 1$):

$$y^1(t,n) = \frac{1}{\sqrt[n]{p(t,n)}} - 1.$$

The one-period risk-free rate used above is computed using simple compounding: $r(t) = y^1(t, 1)$.

The second case is called continuous compounding, where interest is compounded infinitely many times per period ($i = \infty$). Let $y(t, n)$, without the symbol for infinity, denote continuously compounded yields. Fortunately there is a simple formula for continuously compounded yields:[21]

$$y(t,n) = \frac{-\log[p(t,n)]}{n}.$$

In discussing the yield curve, continuously compounded yields will be used. Chart 3 plots the yield curve computed from the discount function that is plotted in Chart 1.

A First Look at the Expectations Hypothesis. The expectations hypothesis can be expressed in a number of equivalent ways. Here is one way to express it: The long-term yield equals the average of the (expected) one-period yields. Of course, when there is no uncertainty, expected one-period yields equal the actual one-period yields. In this case the expectations hypothesis can be expressed as

$$y(t,n) = \frac{y(t,1) + y(t+1,1) + \cdots + y(t+n-1,1)}{n}. \quad (15)$$

But equation (15) is not just a statement of the expectations hypothesis; when there is no uncertainty it is also a statement of the absence of arbitrage opportunities.

That equation (15) is an absence of arbitrage condition when there is no uncertainty can be demonstrated as follows. According to equation (3), the value of an n-period bond today is the

[19] The use of a hedge ratio is analogous to delta hedging in option pricing.

[20] In an appendix to the companion working paper (Fisher 2001), it is shown that the risk premium can be interpreted as a covariance with a marketwide factor. As a consequence, equation (14) has the same form as the capital asset pricing model (CAPM), in which the expected return on an equity equals the risk-free rate plus a risk-premium that depends on the covariance with the market portfolio.

[21] Formally, the continuously compounded yield is the limit of $y^i(t, n)$ as i goes to infinity.

present value of next period's value of an $(n-1)$-period bond:

$$p(t,n) = \frac{p(t+1,n-1)}{1+r(t)} = p(t,1)p(t+1,n-1). \quad (16)$$

The second equality follows from $p(t, 1) = 1/[1 + r(t)]$. Equation (3) can now be applied to the price of an $(n-1)$-period bond at time $t + 1$:

$$\begin{aligned} p(t+1,n-1) &= \frac{p(t+2,n-2)}{1+r(t+1)} \\ &= p(t+1,1)p(t+2,n-2). \end{aligned} \quad (17)$$

Combining equations (16) and (17) produces $p(t, n) = p(t, 1)\, p(t + 1, 1)\, p(t + 2, n - 2)$. This process can be continued until the price of a long-term bond ends up expressed as the product of one-period bond prices:

$$p(t, n) = p(t, 1)\, p(t + 1, 1) \cdots p(t + n - 1, 1). \quad (18)$$

By taking logs of both sides of equation (18) (recall that $\log[ab] = \log[a] + \log[b]$) and dividing by $-n$, equation (15) is obtained.

Since the expectations hypothesis is equivalent to the absence-of-arbitrage conditions when there is no uncertainty, it is understandable that some people may have thought that the same equivalence is true where there is uncertainty—understandable, but wrong.

Uncertainty and Convexity. At this point, the effect of uncertainty on bond yields can be examined. The way in which uncertainty per se drives a wedge between the expected future yields and current yields will be seen. The relation between bond prices and bond yields is not linear; consequently, the yield computed from the average bond price is less than the average yield.[22] In this section, this point is demonstrated and its consequences explored.

The relation between bond yields and bond prices, $y_n = -\log(p_n)/n$, is plotted in Chart 4 for ten- and twenty-year bonds. The two primary features that are evident in the chart are (1) the negative slope and (2) the fact that the graph of the function is "bowed in" toward the origin—in other words, convex to the origin.[23] This second feature is called convexity. Chart 4 shows that a twenty-year bond has more convexity than a ten-year bond.

Convexity drives a wedge between the average yield and the yield of the average price (see Chart 5). There are two outcomes that depend on the flip of the coin: (1) high price and low yield or (2) low price and high yield. The average price and the average yield are at the midpoint of the straight line that connects the two outcomes. But the yield computed from the average price lies on the curved line, below the average yield.

The next step is to derive an algebraic expression for the effect of convexity. Using continuous compounding, the postflip yields can be computed from the two postflip bond prices:

$$y^H_{n-1} = \frac{-\log(p^H_{n-1})}{n-1} \quad \text{and} \quad y^T_{n-1} = \frac{-\log(p^T_{n-1})}{n-1}.$$

The postflip yields can be expressed as

$$y^H_{n-1} = \overline{y}_{n-1} + \delta^y_{n-1} \quad \text{and} \quad y^T_{n-1} = \overline{y}_{n-1} - \delta^y_{n-1},$$

CHART 3 The Zero-Coupon Yield Curve Computed from the Discount Function in Chart 1

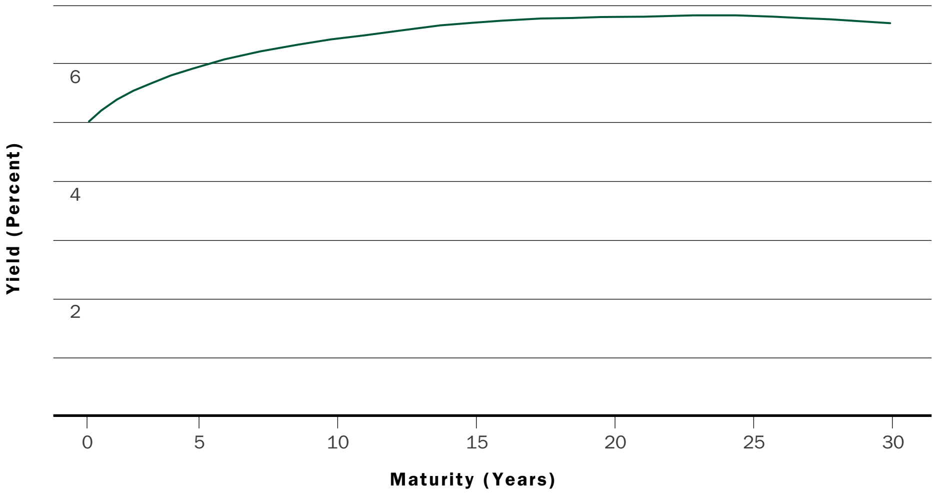


CHART 4 The Yield of Zero-Coupon Bonds as a Function of the Price

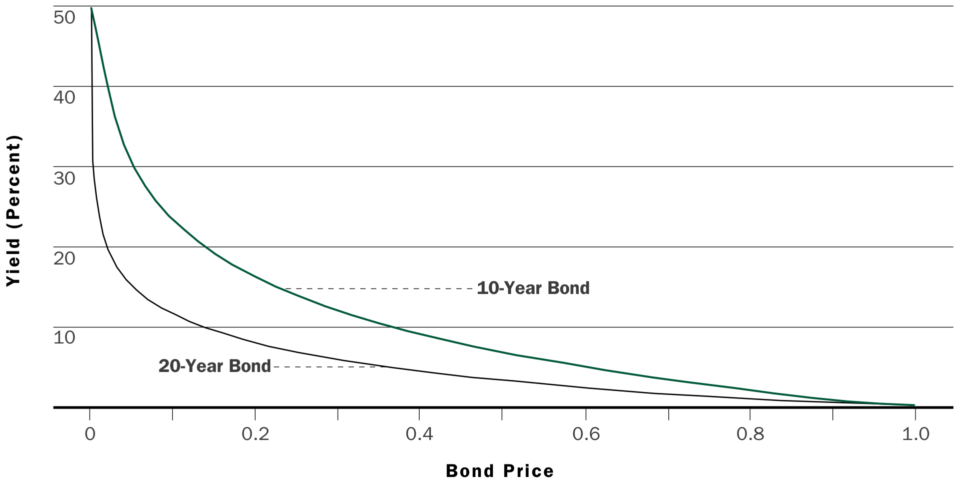
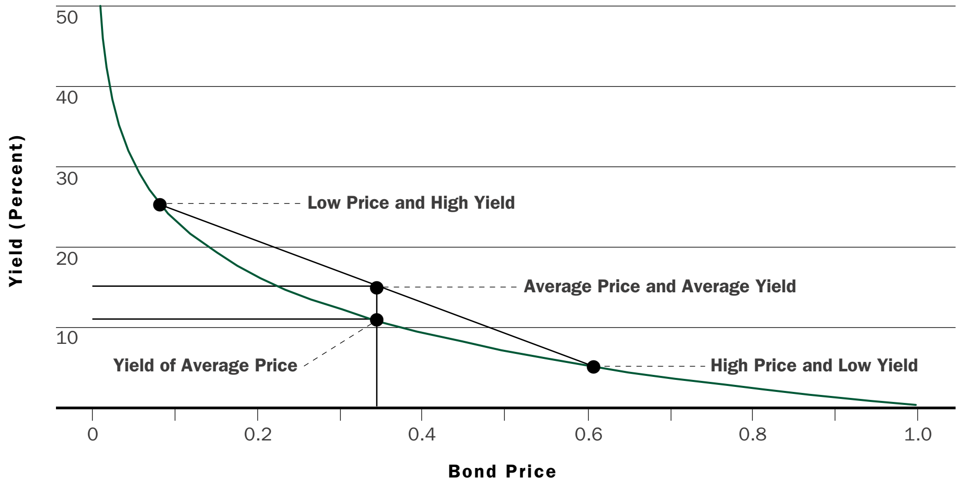


CHART 5 The Convexity Effect on the Average Yield and the Yield of the Average Price



where the average yield and the volatility of the yield are given by

$$\bar{y}_{n-1} = \frac{y^H_{n-1} + y^T_{n-1}}{2} \quad \text{and} \quad \delta^y_{n-1} = \frac{y^H_{n-1} - y^T_{n-1}}{2}.$$

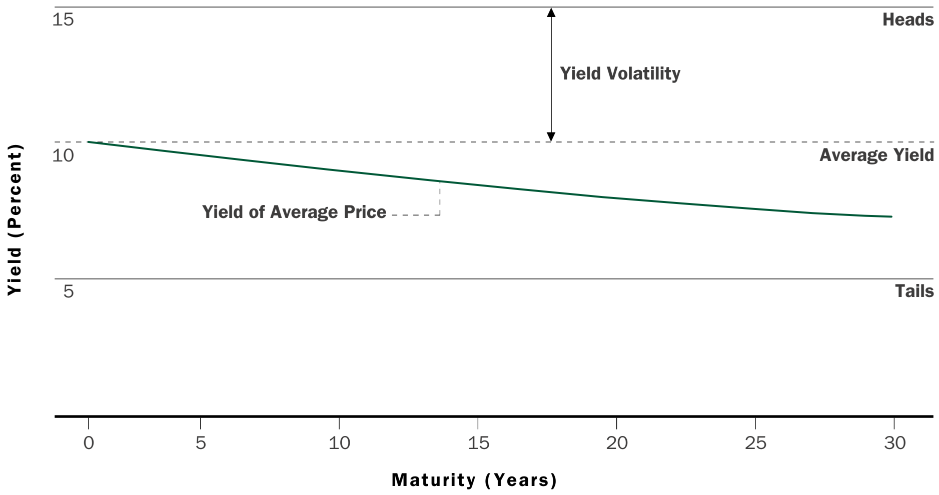
As an example, suppose that the current one-period yield equals the average long-term yield, $y_1 = \bar{y}_{n-1} = \bar{y}$, for all $n \geq 2$, and also suppose that the yield volatility is constant, $\delta^y_{n-1} = \delta^y$. Then the yield on an n-period bond (that is, the yield curve) can be approximated by $y_n \approx \bar{y} - n(1/2)(\delta^y)^2$ as long as n is not too big. This approximation illustrates the three main features of convexity: (1) Convexity has the effect of reducing yields. (2) The convexity effect is larger for longer-term bonds. (3) The convexity effect depends on the variance of the uncertainty about yields. See Chart 6 for an example in which $\bar{y} = 0.10$ and $\delta^y = 0.05$.[24]

[22] This result is an example of Jensen's inequality.

[23] These two features summarize the first two derivatives of the bond yield with respect to the price. The first derivative is negative, and the second derivative is positive.

[24] The graph is drawn using the exact formula upon which the approximation is based. See Part 2 of the companion working paper (Fisher 2001) for the details.

CHART 6 An Example of the Convexity Effect on Zero-Coupon Yield Curves



Note: Where $\bar{y} = 0.10$ and $\delta^y = 0.05$

This example illustrates the depressing effect of uncertainty on bond yields via the convexity effect. As noted in the introductory section, risk premia will also have an effect on the shape of the term structure. Unfortunately, it is beyond the scope of this article to treat the effect of risk premia in a fully rigorous way. (The interested reader will find a full account in Part 2 of the companion working paper [Fisher 2001]).

Conclusion

The conditions that ensure the absence of arbitrage opportunities provide structure for the analysis of the yield curve by appealing to rationality at its most basic level. The central implication of the no-arbitrage conditions is that the risk premium for an asset can be decomposed into the amount of risk (measured by volatility) and the price of risk (which reflects investors' attitudes toward risk), where the price of risk is common to all assets. The forces that shape the yield curve are channeled through this feature. The nonlinear relation between bond yields and bond prices leads to surprising and even counterintuitive results. It is necessary to have a firm grasp of the no-arbitrage conditions in order to make sense of the shape of the yield curve. Analysis that ignores the implications of the no-arbitrage conditions will inevitably lead one astray.

REFERENCES

Black, Fischer, and Myron Scholes. 1973. The pricing of options and corporate liabilities. *Journal of Political Economy* 81 (May/June): 637–54.

Cox, John C., Jonathan E. Ingersoll Jr., and Stephen A. Ross. 1981. A reexamination of traditional hypotheses about the term structure of interest rates. *Journal of Finance* 36 (September): 769–99.

Fisher, Mark. 2001. Forces that shape the yield curve: Parts 1 and 2. Federal Reserve Bank of Atlanta Working Paper 2001–3, March.

Vasicek, Oldrich. 1977. An equilibrium characterization of the term structure. *Journal of Financial Economics* 5, no. 2:177–88.

APPENDIX

Notations

Bond-Price Uncertainty

p_n	Value of an n-period bond (same as $p[t, n]$)
r	One-period interest rate (same as $r[t]$)
p^H_{n-1}	Value next period of an $(n - 1)$-period bond if the coin comes up heads
p^T_{n-1}	Value next period of an $(n - 1)$-period bond if the coin comes up tails
$\bar{p}_{n-1}$	Average value of an $(n - 1)$-period bond (preflip)
δ^P_{n-1}	Volatility of the bond price (amount of risk)
a_{n-1}	Adjustment term (risk premium)

Bond Portfolios

b	Number of m-period bonds held in portfolio
b^*	Number of m-period bonds held to make the portfolio risk-free
π^H	Value of the portfolio next period if the coin comes up heads
π^T	Value of the portfolio next period if the coin comes up tails
π^*	Value of a risk-free portfolio (holding b^* m-period bonds)
λ	Price of risk

Bond Yields and Compounding

$y^i(t, n)$	Yield at time t on an n-period bond, compounded i times per period
$y^1(t, n)$	Yield computed with simple compounding
$y^1(t, 1)$	Same as $r(t)$
$y(t, n)$	Continuously compounded yield (same as $y^\infty[t, n]$)

Bond Yield Uncertainty (Continuously Compounded)

y_n	Yield at time t on an n-period bond (same as $y[t, n]$)
y^H_{n-1}	Yield next period on an $(n - 1)$-period bond if the coin comes up heads
y^T_{n-1}	Yield next period on an $(n - 1)$-period bond if the coin comes up tails
$\bar{y}_{n-1}$	Average yield on an $(n - 1)$-period bond (preflip)
δ^y_{n-1}	Volatility of the yield

Is Why We Use Money Important?

VICTOR E. LI

Li is a senior economist in the macropolicy section of the Atlanta Fed's research department. He thanks Karsten Jeske and Ellis Tallman for helpful comments.

MONEY PLAYS A CENTRAL ROLE IN DETERMINING THE COURSE OF MACROECONOMIC ACTIVITY. PRICES AND INFLATION ARE DIRECTLY LINKED TO THE NATION'S MONEY SUPPLY, AND MANY ECONOMISTS BELIEVE THAT CHANGES IN THE QUANTITY OF MONEY ALSO HAVE IMPORTANT EFFECTS ON REAL ECONOMIC VARIABLES, SUCH AS UNEMPLOYMENT AND gross domestic product, especially in the short run. Yet many of the models that economists use to evaluate fundamental questions relating money and monetary policy to economic activity tend to gloss over the underlying characteristics of the economy that motivate the use of money. Economic models that simply assume currency is valued overlook these characteristics and possibly the important properties of money that influence the way its supply affects the economy. Understanding these properties will provide a better idea of not only the key features of money that associate it with "value" but also how those characteristics affect the link between the quantity of money and aggregate economic activity.

Economists define money by its functions as a medium of exchange, a store of value, and a unit of account. Its function as a medium of exchange is unique.[1] Money serves as an alternative to interest-bearing assets because it is the easiest asset to exchange directly for goods and services; that is, it provides liquidity services. Macroeconomic models often assume exogenously that money is used in carrying out transactions. One approach is simply to specify a money-demand relationship that says the demand for currency depends positively on income and negatively on the nominal interest rate (a relationship that empirical money-demand studies have confirmed). Another widely used approach is to capture the notion of money demand within the context of neoclassical economic models. Such models base macroeconomic outcomes on the microeconomic decisions of households and firms in perfectly competitive markets. These approaches typically motivate the use of money in transactions in one of two ways. Cash-in-advance models impose a constraint that says current expenditures must be financed with previously accumulated holdings of cash. Money-in-the-utility-function models treat money as a special asset that yields satisfaction, or utility, to the holder. The idea behind this approach is that money's liquidity services, not money itself, provide utility to individuals.

At first glance, it seems very reasonable to simply assume that money is used to buy goods and services. After all, this assumption is true in virtually all modern economies. Yet it presents some fundamental problems, especially within the context of neoclassical macroeconomic models. The hallmark of neoclassical economics is the idea that markets are perfectly competitive and that no difficulty exists in trading goods for goods. Goods and services are exchanged in a centralized marketplace, and an auctioneer coordinates trades and ensures that they occur at market-clearing prices. In such a framework, there is no need for a medium of exchange and no role for money. Simply assuming a cash-in-advance constraint or that the liquidity services of money provide utility is thus inconsistent with the underlying economic environment of the neoclassical model. In order to explain why individuals use money, there must be frictions in the transactions process that make trade difficult.

These trade frictions that underlie the value of money may be crucial in addressing two types of fundamental questions in macroeconomics. The first type deals with how monetary policy affects economic activity: What effect do changes in the growth rate of the money supply have on productive activity? What are the consequences of inflation on economic welfare? Economic models that ignore the trade frictions that explain money's usefulness inevitably overlook how changes in economic policies affect those frictions. For example, monetary policies may be used to reduce trade frictions by making exchange easier and encouraging buyer and seller participation in the market. This effect would be overlooked in economic models that do not account for the frictions that cause money to be used in the first place.

A challenge to monetary economics is to construct models that explicitly capture the transactions role of money and can be used to address central issues in monetary economics and macroeconomics.

The second set of questions deals with the choice and use of alternative currencies in transactions: In particular, why would different (international) currencies circulate within a particular country, and why might a group of countries adopt a common currency? It is not possible even to address such questions in economic models that assume at the outset that a particular currency is used. The challenge to monetary economics is to confront these fundamental problems by constructing models that explicitly capture the transactions role of money and can be used to address central issues in monetary economics and macroeconomics.

A promising class of models that responds to this challenge is the search-theoretic approach to money. At the heart of this approach is the idea that because transactions take place at different places and times, meetings between buyers and sellers are not instantaneous. Individuals must spend time or resources to search for sellers who have goods they would like to consume and who are willing to trade them for something the buyer possesses. In such an environment barter is costly because of the difficulty of finding a "double coincidence of wants." That is, for barter to take place, not only must the buyer want the good the seller has, but that seller must also want the good the buyer has. A universally acceptable medium of exchange, which may be intrinsically useless fiat money, is valued because it overcomes the double-coincidence-of-wants problem associated with barter, thus making trading easier. This function of money is certainly not new; it dates back to the earliest writings of classical economists more than a century ago (Jevons 1875; Wicksell 1911). Yet until recently it had not been well formalized using the tools of economic theory.

This article will first explain how search and matching models of money identify the characteristic assumptions for motivating the use of money in carrying out transactions.[2] The pioneering work in this area by Kiyotaki and Wright (1989, 1991, 1993) was driven by some fundamental questions in monetary economics: Under what conditions will objects emerge that circulate as mediums of exchange? Which objects, including commodities, would circulate, and how important are their intrinsic properties and the extrinsic beliefs regarding their acceptability? How precisely does the use of money as a medium of exchange affect liquidity and welfare?

While the search-theoretic approach seems well-suited for studying such abstract and intellectual questions in monetary theory, one wonders how such a model can be useful to modern economies. This article uses two approaches to argue that search models of money do indeed shed light on some important issues in modern macroeconomics that may have empirical relevance. First, because search models make internal, rather than ex ante, assumptions about which currencies emerge as mediums of exchange, such models are uniquely qualified to study questions that arise naturally in international monetary economics. For example, what determines which currencies are used in a particular country, and what are the costs and benefits to a country of having its currency circulate internationally? These questions are relevant to "dollarization" issues or, more generally, currency substitution. Second, an explicit treatment of the transaction role for money may have significant implications for the economic impact of monetary policy and inflation. For example, an increase in the money-supply growth rate and inflation can provide an incentive for buyers and sellers to participate more actively in the market. In turn, a higher degree of market participation can affect the ease of finding trading partners, hence reducing the severity of trade frictions in the marketplace. This additional channel by which monetary expansions can affect economic activity may be important in stable-price, low-inflation countries. It also provides an alternative to the Keynesian sticky wage and price literature, which finds a positive link between money, inflation, and output.

This article summarizes some of the recent literature on search models of money and their successful application to these issues. The article is not intended to provide a comprehensive review of the search-money literature. Instead, it highlights representative contributions that demonstrate the applicability of search models to monetary and macroeconomic theory and suggests directions for future work.

When considering search-theoretic approaches to money, one should keep in mind that economic theory inevitably involves a high degree of abstraction, especially in macroeconomic models based on microeconomic behavior. The models economists use are vast simplifications of the real world. This stylized world helps isolate and analyze the relationship between a few important variables of interest. The assumptions these models make and the way they capture interactions between economic participants are necessarily approximations that will not accurately represent all aspects of the actual macroeconomy. Yet economic models can be extremely useful in two related respects. First, they provide a logically consistent framework for analyzing the interrelationships between important economic variables. Second, if the outcomes of the models capture important features observed in reality, then these models can have predictive power.

What Is a Search Model of Money?

Search-theoretic models focus precisely on the various frictions motivating the use of money as a medium of exchange. These frictions are characterized by the following properties: (1) a separation of market participants: there is no centralized marketplace, and all trades do not occur at the same place at the same time; (2) differentiated goods: there are many different types of goods and many individuals with different tastes; and (3) anonymous trading with no public "legacy:" trades occur anonymously in that the trading histories of each individual are not public information.

The first two properties lead to a double coincidence problem with barter. Ms. Burger Queen, who makes hamburgers but likes pizzas, must find Mr. Pizza Delight, who makes pizzas and likes hamburgers. This search takes time because many other individuals in the economy produce and consume different types of goods. For example, as Ms. Burger Queen searches for an opportunity to trade her hamburgers for pizzas, she might encounter Mr. Pizza Express, who makes pizzas but needs a haircut. Since there is no double coincidence of wants, only a single coincidence, trade will not take place. This simplistic story illustrates that locating a barter exchange is a time- and resource-consuming process that makes trade complicated without a medium of exchange.

The third property, anonymous trading, permits money to act as an objective record keeper of past actions and enables otherwise impossible transactions. For example, suppose that Ms. Burger Queen, who makes burgers and likes pizzas, meets Mr. Pizza Express, who needs a haircut. Ms. Burger Queen may request that Mr. Pizza Express give up a pizza (in exchange for nothing, since she does not know how to give a haircut) because she previously gave up a hamburger to another individual for nothing and would like compensation for her good deed or because in the future she promises to respond in kind to another individual. If everyone were willing to follow through on his or her promise to respond in kind then this "credit arrangement," involving giving up goods in exchange for nothing but a promise that others will respond in kind, would be the most efficient means of exchange. However, if Ms. Burger Queen's trading history is not public information or there were no way to enforce her commitment to respond in kind in the future—that is, if trade is anonymous—then Mr. Pizza Express would have no incentive to enter into this arrangement.

In anonymous trading, when people trade with others they do not know, they are not willing to engage in any exchange that is not quid pro quo. A medium of exchange substitutes for an abstract promise to respond in kind and hence acts as a record keeper in a world with anonymous trading. Search models capture anonymous trading by

[1] Because a medium of exchange must be held during the time after income is received and before it is used to purchase goods and services, it is also a store of value. Stores of value, such as interest-bearing financial assets, do not necessarily serve as mediums of exchange, however.

[2] Search models are sometimes called matching models because of the way buyers and sellers meet or "match" with each other in the market.

assuming individuals are matched with each other randomly and one-to-one for the purpose of trade. These models of money show explicitly how these three properties lead to the use of money as a medium of exchange.

A Prototype Search Model of Fiat Money. Kiyotaki and Wright (1991, 1993) formalize the way acceptability ultimately drives the use of fiat money as the medium of exchange in an economy with the type of trade frictions described above. The intrinsic properties of commodity monies, such as gold or silver objects, may make them more acceptable than others. Yet acceptability lies at the heart of a social choice of a medium of exchange without intrinsic value. Consequently, both of these seminal studies focus on fiat rather than commodity money.[3] To demonstrate these concepts, this article outlines the important elements of a prototypical search model of money and discusses its findings. The assumptions of the model are very simplistic and will not accurately reflect various aspects of real economies. The model is designed to illustrate how intrinsically useless fiat money can become an acceptable medium of exchange, that is, the model investigates the circumstances that determine when fiat money will be valued as an equilibrium outcome. Furthermore, the logic behind the model's main ideas will also hold in more complex economic environments that come closer to reality.

Imagine an economy with many different types of goods and individuals. Individuals have different tastes and therefore would like to consume only a small fraction of the total goods. Call this fraction x. For example, there could be three types of goods and three types of people in the economy: fruit lovers who like only fruit, vegetarians who consume only vegetables, and carnivores who like only meat. If each individual likes only one-third of the overall goods produced in the economy, $x = 1/3$. Hence, x is a measure of the overall acceptability of goods in the economy. Individuals specialize in the production of a particular type of good and look for opportunities to trade for the goods they would like to consume. In this example, a vegetable producer may like to eat only fruits. Also assume that individuals are able to produce and carry only one unit of a good at a time as they search for opportunities to trade. Thus, if an exchange does occur, it will be a one-for-one swap of goods for goods. Finally, individuals value their time and would prefer to consume sooner rather than later.

The model is dynamic in that it considers the behavior of these individuals over time. At the beginning of the "day," each individual produces one unit of her good (her "production good") and enters the market to search for opportunities to trade her good for one she likes to consume (her "consumption good"). For trade to occur, each individual must find someone with a double coincidence of wants. Recall that x is the fraction of all goods in the economy that each individual finds desirable. As each person (randomly) meets others in the market, the chance that she will like the good another individual is carrying will be x, and the chance that another individual likes her good will also be x. Thus, the probability of a double coincidence of wants for each meeting is $(x)(x) = x^2$. If x is small, say 1/3, then the probability of a double coincidence of wants will be much smaller, $x^2 = (1/3)(1/3) = 1/9$. This comparison is what makes barter difficult in the model. Every time an individual meets a potential trading partner, there is only x^2 of a chance that the individuals will actually be able to trade for the goods they want to consume. It takes time to find that double coincidence of wants, and individuals value time. In this barter economy, once a double coincidence of wants is found, trade occurs, the individual enjoys consumption, she produces another good, and the process begins again.

Next consider the introduction of fiat money into this economy—say, pieces of paper that have no intrinsic value. To keep everything simple, assume that individuals can each hold at most one unit of money or one unit of a good that they produce themselves (but not both). These units of money (dollars) are also indivisible (that is, they cannot be divided into quarters, dimes, and so on). Let M denote the fraction of the population holding money and $(1 - M)$ the fraction holding goods. Since individuals are either holding one unit of money or none, M also corresponds to the total money supply. Additionally, these assumptions imply that an exchange is simply a one-for-one swap of goods for goods or goods for money. Each individual believes that if she holds money, the chances that a seller will accept that money for her production good is Π. If $\Pi = 1$ then she believes that all individuals are willing to accept money for her good, if $\Pi = 0$ then she believes no one will accept money for her good, and if $0 < \Pi < 1$ she believes money may be accepted sometimes. Thus, Π represents the economywide acceptability of money.

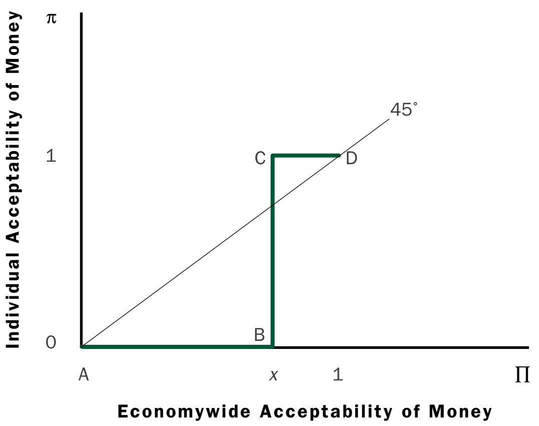
Suppose an individual begins the day with a good that he has produced himself and proceeds to the market to look for trading opportunities. The chance that he will encounter another individual also holding a good is now $(1 - M)$, the fraction of the population holding goods. But, as before, he will be able to barter only if there is a double coincidence of wants, and the chances of that happening may be very small (x^2). On the other hand, the probability that he will encounter individuals holding money is given by M. The chance that these individuals would like to buy his production good with money is given by x. When this happens, he as the seller must decide whether or not he is willing to surrender his production good for this intrinsically valueless piece of paper. This "willingness" is denoted as π. The possible choices for π are

$\pi = 1 \quad \Rightarrow$ he always accepts the money;
$\pi = 0 \quad \Rightarrow$ he never accepts the money.

The seller's choice will depend upon his belief in the "economywide willingness" of other people to accept money in exchange for their goods, defined earlier as Π. If another person likes the seller's goods and he is willing to accept money, he will trade his good for money. If this exchange happens, the seller now becomes a buyer with a unit of money and must search for an individual who has a particular good, an event with the probability ($1 - M$). The chance that another individual has his consumption good is x, and the probability that the other individual is willing to accept money in exchange for that good is Π. If the buyer can exchange his unit of money for his consumption good, he consumes it and produces another unit of his production good, and then the process begins again.

The key question is, under what conditions would an individual be willing to surrender a production good for an intrinsically useless piece of fiat money, given the individual's beliefs about Π? The answer can be summarized by a simple diagram, illustrated in the chart. Along the horizontal axis is Π, which can be viewed as the independent variable since each individual takes his belief about the economywide acceptability of money as given when deciding whether to accept money. This decision to accept money is the dependent variable π, plotted along the vertical axis. The darker line connecting points A, B, C, and D shows the value of π that individuals who are acting in their own best interests should choose for a given Π. The individual choice of whether or not to accept money, π, will ultimately determine the economywide, or equilibrium, value of Π.[4] The 45 degree line indicates that because everyone behaves in a similar way, an equilibrium is a situation where $\pi = \Pi$.

Equilibrium in the Prototype Search Model of Money



Hence, an equilibrium in the model economy is a situation where the 45 degree line intersects with the line ABCD.

Suppose individuals believe that Π = 0; that is, no one else in the economy is willing to trade his or her goods for money. Giving up one's own good for money would then be foolish because the money would be essentially useless. Hence, the best decision is never to accept money and to choose π = 0 as well. In fact, as long as $\Pi < x$, so that money is less acceptable in exchange than goods, individuals should still choose π = 0 and never accept money. This outcome is shown as line AB along the Π-axis between 0 and x. Given $\Pi < x$, there is a *pure barter equilibrium*, in which no one accepts money in exchange for goods and π = Π = 0. This event is shown where line AB intersects with the 45 degree line at point A.

Now consider the other case where Π = 1. In this case, individuals believe that all others in the economy are always willing to give up their production goods in exchange for money. Consequently, each individual should be willing to do the same. Why? If someone does not trade her good for money, the chance that she will be able to barter with someone holding her consumption good is the double-coincidence probability of x^2. By accepting cash today, however, the chance that she will be able to trade that money for goods is merely the likelihood that she likes another's good, or x, which is greater

[3] Kiyotaki and Wright (1989) use a search-theoretic model to study the properties of commodity monies that lead them to be used as mediums of exchange.

[4] Economists call this equilibrium concept a Nash equilibrium from game theory. Everyone is pursuing the best strategy he or she can, given the actions of others. The situation is a symmetric Nash equilibrium if all individuals end up pursuing identical strategies.

than x^2. That is, because money is always acceptable to others, all that is needed is a single coincidence of wants to trade money for goods instead of a double coincidence of wants. If everyone else accepts money, then it will be in each person's best interest to do so as well and to set $\pi = 1$.

Following the same logic, as long as $\Pi > x$, money is more acceptable in exchange than goods, and individuals should still choose $\pi = 1$ and always accept money themselves. The line CD at $\pi = 1$ between x and 1 demonstrates this concept. Therefore, if $\Pi > x$, then there is a *pure monetary equilibrium*, in which everyone accepts money in exchange for goods and $\pi = \Pi = 1$. Graphically, this equilibrium is where the 45 degree line intersects line CD at point D. These results demonstrate that monetary exchange is a self-fulfilling prophecy. Money is accepted in exchange simply because of the belief that others will accept it as well.[5]

Because search models make internal, rather than ex ante, assumptions about which currencies emerge as mediums of exchange, such models are uniquely qualified to study questions that arise naturally in international monetary economics.

Kiyotaki and Wright next analyze how the use of money affects social welfare or the overall lifetime welfare of individuals participating in their model economy. They show that if finding a double coincidence of wants is very difficult—that is, if x is small enough—then the use of money is socially beneficial. It minimizes the search costs associated with exchange by providing liquidity. Increasing the money supply, M, however, does not always increase liquidity. Because money in this model is indivisible and individuals can hold at most only one unit, an increase in the money stock, M, necessarily increases the number of people holding money and reduces the number of people holding goods, $1 - M$. Thus, increasing the money stock "crowds out" goods. Hence, too much money may also make trading difficult by reducing the number of sellers with goods in the market ("too much money chasing too few goods"). This relationship can be seen in the quantity exchange equation, in which the velocity of money (V) is given by the ratio of nominal output to the number of dollars in circulation, or $V = PY/M$, where P is the price level and Y is real output. Since individuals can carry only a unit of cash or goods, total output is just the fraction of those individuals not holding money, or $Y = (1 - M)$. Furthermore, one-for-one exchanges of goods and money imply a nominal price of one ($P = 1$). Thus, $V = (1 - M)/M$. Increasing M will always reduce the velocity of money. This reduction is reenforced by the crowding out of goods. Eventually, increases in the money supply will inhibit liquidity if M is high enough.[6]

What about Prices? The prototypical search model of money outlined above provides a logically coherent theory of money as a medium of exchange. Yet it also relies on some very special assumptions. Among them is the notion that goods and money are indivisible and all exchanges are one-for-one swaps of goods and money. This notion necessarily implies that nominal prices are exogenously fixed and equal to one. It is only a first step in thinking about money as a medium of exchange. While the model captures the protocol of exchange, it says nothing about the determination of the prices at which exchanges occur.

A straightforward way around this difficulty, explored by Trejos and Wright (1993, 1995) and Shi (1995), is to incorporate *bilateral bargaining* as a mechanism to determine prices. The idea is to think of goods as exchangeable services that are divisible but not storable. Examples of these types of goods are haircuts and perishable fruit. Hence, when individuals meet, these goods must be produced and immediately traded. When a buyer and seller meet, they bargain over the price that divides the gains from trade between them in a particular way. A nominal price would simply be the ratio of the quantity of goods, q, exchanged for a unit of indivisible currency, $p = 1/q$. This price is affected by factors that influence an individual's bargaining power and opportunities to walk away from a trade. Such factors include the aggregate quantity of money and the ease of finding other trading partners. This model is the prototype search model of money with prices.

Under certain conditions, an increase in the money stock, M, can lead to lower prices if M is small and higher prices if M sufficiently large. If M is very low, then an increase in the money supply enhances market liquidity and the bargaining power of buyers. In order to implement the trade, sellers must reduce the asking price on their goods. Just as in the prototype model, however, high values of M can impede liquidity by reducing the quantity of goods available for trade. This reduction distributes bargaining power in favor of sellers, and prices rise with increases in the money stock. This (imperfect) link between the quantity of money and liquidity also has other real effects. For example, if M is low, the frequency of transactions, output, and welfare can increase with a higher money stock. The eventual "inflationary" effects of continuously increasing M, however, lead to a decline in these variables. Thus, the quantity of money can have important effects on output by influencing the relative difficulty of how buyers and sellers meet and the prices negotiated at those trades. Monetary policies should not overlook such liquidity effects when an "optimal quantity of money" is being formulated.

These results demonstrate that prototype search models of money make a significant contribution to the pure theory of money. By explicitly considering the frictions that make trade difficult, they demonstrate the necessary conditions that lead individuals to use money rationally as a medium of exchange. These models can also be used to study, in a stylized fashion, the links between the quantity of money, liquidity, and prices. Such approaches, however, are by no means limited to the pure theory of money. Next, this article discusses these models' usefulness in examining important monetary and macroeconomic issues.

International Currency

When a particular country's currency is also used for transactions in other countries, it is called an international currency. Issues involving international currency are at the forefront of today's economic headlines. For example, what are the long-term economic consequences of European countries' adopting the euro? Should Latin American countries abandon local currencies in favor of the dollar? One of the distinguishing features of search models of money is that they are designed to examine which currencies are used to carry out transactions. Traditional international models that simply assume the use of a particular currency have nothing to say about these issues. Search-theoretic approaches emphasize the endogenous determination of mediums of exchange and are naturally suited to study the use of dual and international currencies.

Matsuyama, Kiyotaki, and Matsui (1993) take the first step in applying search-theoretic models to study international currencies. They consider two countries, each of which issues its own currency. The populations of both countries are growing, and each government issues its money through purchasing goods from individuals—that is, money creation generates seigniorage revenue. Individuals are randomly matched for the purpose of trade, with two individuals from the same country meeting more frequently than individuals from different countries. Even with the assumption that exchanges are one-for-one swaps of goods and money and involve fixed nominal prices, Matsuyama, Kiyotaki, and Matsui are able to analyze such issues as how and when local currencies can survive in the presence of a universally accepted international currency and whether an international currency emerges naturally as economies become more integrated.

Generally, several types of outcomes are possible: there may be no international currencies, one country's currency may circulate in both countries, or both may circulate in both countries. The authors find that the degree of economic openness, as measured by how often buyers and sellers from each country interact, is central in determining which currencies are acceptable in trades. For example, if it is relatively easy for Mexican buyers to meet Mexican sellers but relatively hard for U.S. buyers to meet Mexican sellers, then pesos will not circulate in the United States. In addition, as long as the rate at which Mexican buyers meet U.S. sellers is not too low, dollars will circulate as the international currency in both countries. Such outcomes are possible if the supply of international currency is not too abundant and the noninternational currency is not too scarce.

Incorporating prices into the framework via bargaining allows the model to address a host of additional issues regarding purchasing power and exchange rates. How, for example, does the international circulation of a currency affect its purchasing power at home? How are policies designed to achieve higher levels of seigniorage or welfare affected by currency substitution? Trejos and Wright (1996, 2000) explore these issues. A variety of outcomes is possible, and conditions similar to those in Matsuyama, Kiyotaki, and Matsui (1993) must hold for one international currency to exist. In this case, Trejos and Wright they find that the international currency will have more purchasing power at home than abroad, and it will have more purchasing power in the foreign country than the noninternational currency.

This result also has an interesting implication for the observed empirical failure of purchasing power parity. Many empirical studies find that when prices are converted to a common currency, richer countries tend to have higher prices. While there are various explanations of this phenomena, none

[5] There is an intermediate case in which $\Pi = x$. In this case, money and goods are equally acceptable in exchange, $\pi = \Pi = x$, and the equilibrium occurs where the 45 degree line intersects line BC.

[6] In such a situation, although an individual seller can easily trade his or her good for money, economywide liquidity is inhibited since almost all individuals in the economy are buyers with money who cannot find sellers with goods.

explain why the United States is an exception to this regularity. That is, U.S. prices are much lower than predicted by cross-country income-price level regressions (Balassa 1964; Rogoff 1996). Search models offer a unique perspective on this issue. If the dollar is used in other countries as an international currency, the model predicts that its value in the United States should be higher, implying a lower domestic price level. Intuitively, acceptability and liquidity give money value in this model, and acceptability and liquidity abroad enhance its value at home.

This framework also provides some novel policy implications. For example, if both the domestic and a foreign currency circulate within a country, then maximizing seigniorage would require the domestic money supply to exceed its welfare-maximizing level. This requirement is so because dual currencies diminish the degree to which domestic currency is used, hence also diminishing the "inflation tax base." Increased coordination between the monetary policies of the two countries, however, may actually imply that increases in seigniorage and welfare are consistent with lower money supplies in both countries relative to noncooperation.

Search models have been extended to study other related issues regarding the use of multiple currencies. Zhou (1997) allows the possibility of currency exchanges in the model outlined above; Craig and Waller (1999) study the impact of currency reform and apply their framework to the recent effort of the Ukrainian government to remove the U.S. dollar from its economy and encourage the use of a new domestic currency; and Curtis and Waller (2000) consider illegal and black market currency exchanges.

Money, Inflation, and Economic Activity

One of the most prominent and debated issues in macroeconomics is the impact of the quantity and growth rate of money on inflation and economic activity. Earlier Keynesian macroeconomic models based on rigidities in prices and wages predicted a positive trade-off between inflation and productive activity. New classical models predict that while monetary policy can have important short-run effects, such as when changes in the money supply are unanticipated, inflation in the long run will be detrimental to economic activity. These new classical findings also hold in models that approximate the transactions role of money via a cash-in-advance constraint or placing money into the utility function. In these models, higher money growth creates inflation, taxing all activities involving cash. Individuals reduce their cash holdings, buy fewer goods and services, and overall economic activity declines. Yet all of these approaches simply assume an exogenous transactions role for money, overlooking the ways monetary policy and inflation can affect the very frictions that cause money to be used.

Short-Run and Long-Run Effects of Money. Traditional macroeconomic theory suggests that there are important differences between the short-run and long-run effects of monetary policy (Friedman 1968; Lucas 1973; Barro 1976). In particular, the common belief is that in the short run an increase in the money supply tends to expand output, while the long-run effect is primarily price inflation. Wallace (1997) investigates whether these results logically proceed from a matching model of money. His framework builds on the prototype search model of money with prices and considers the effect of a one-time increase in the stock of money, M. Since money is indivisible and individuals hold at most one unit, only those not holding cash may receive the cash transfer. The model incorporates a special assumption regarding what individuals know about the size of the increase in M. In the period in which M increases, or the short run, the public knows new currency has been distributed to a portion of the population, but does not know how much has been distributed. An individual who receives one unit of cash today does not learn how many others have also received cash. In the period following the short run, all individuals learn about the size of the increase. This period and all subsequent periods are called the long run.

Wallace finds that in the short run, increasing the money supply in the above manner leads to an increase in output with no impact on the price level. Intuitively, prices are already determined based upon the expected size of the increase in M. Since individuals do not know the actual size of this increase, it has no effect on current prices. The additional money in the economy enhances liquidity, which in turn increases the frequency of transactions and output. Once individuals discover the size of the increase in M in the long run, bargaining between buyers and sellers results in a higher price level much the same as in Trejos and Wright (1995). While the frequency of transactions is higher, the quantities exchanged are lower, leading to ambiguous effects on overall output. Thus, a search model of money replicates a macroeconomic feature of money most economists would agree with: while short-run changes in money have real effects, long-run changes have primarily nominal effects. The model also conforms with the view that information lags are important in explaining the real effects of unanticipated changes in monetary policy.

How Can Search Models of Money Study Inflation? The prototype search model of money with prices provides a first step in understanding how such models can be used to study the interaction between money, prices, and economic activity. Their applicability is severely limited, however. Recall that these models assume money comes in indivisible units and cannot be divided. Since individuals can hold only one unit of money or a good, the quantity of money is directly linked to the proportions of the population holding either money, M, or goods, $(1 - M)$, and increasing the money supply necessarily crowds out goods. Furthermore, these assumptions make it impossible to analyze policies that change the growth rate of the money supply and the inflation rate.

Three approaches have been used in the literature to deal with these issues. First, and perhaps most straightforwardly, one could use the prototype model and approximate the effects of inflation by a tax on fiat money (Li 1994, 1995). However, since there are essentially no prices in the prototype search model of money (they are fixed at one) and the money stock is fixed, this approach only suggests how search frictions affect the interaction between inflation and welfare.

To capture the notion of money growth in a more realistic way, these search models must be able to account for the individual decision to demand and hold many units of currency at a given moment. The second approach uses this methodology by directly generalizing the prototype search model and allowing individuals to accumulate many units of cash (Molico 1999; Camera and Corbae 1999). A third approach is to have individuals choose divisible money holdings in an environment in which they are interacting with a very large number of buyers and sellers over the shopping period (Shi 1997, 1999; Laing, Li, and Wang 1999, 2000). Each of these approaches has its advantages and has proved useful in studying various aspects of the money supply and the inflation process. The following subsections will discuss each approach in more detail.

Inflation as a Tax on Money. In the prototype search model of money, one can draw some implications of inflation and welfare by thinking about the effects of taxing fiat money. Such a methodology allows one to retain the simplifying assumptions that goods and money are indivisible and exchanges are one-for-one swaps. Li (1994, 1995) pursues this methodology in two articles. First, these studies modify the model so that, instead of a completely random matching process, individuals can choose the frequency with which they contact others in the market. In such an environment, individuals tend to invest too little effort in search, and the frequency of transactions is too small to be socially beneficial. Individuals consider only their own private gains, rather than social gains, when deciding how much to search.

The studies then consider the effects of imposing a tax on money balances in a way that resembles an inflation tax. In this process, the government obtains seigniorage revenue by randomly confiscating money from money holders and using it to purchase goods from goods holders. The similarity between this taxation rate and actual inflation is that as a buyer shops with money, there is a chance that his money holdings will be confiscated and devalued (made worthless). The studies show that such a taxation process can increase the level of individual investment in search. Intuitively, if the chances of having one's money taxed increase, the incentives to find and exchange one's money for a desirable consumption good increase as well.

The intrinsic properties of commodity monies may make them more acceptable than others. Yet acceptability lies at the heart of a social choice of a medium of excchange without intrinsic value.

This relationship improves the overall rate of transactions and may lead to an improvement in economic welfare. Moreover, if individuals are able to accumulate inventories of goods, as in Li (1994), they may hedge against the inflation tax by increasing their inventory stocks. This result is analogous to individuals shifting holdings out of money and into nonmonetary assets when inflation is high. The findings suggest that a low but positive inflation tax may actually have beneficial effects by promoting inventory accumulation and the frequency of transactions. This measure, precluded by monetary models that ignore trade frictions, would tend to counter the costs of inflation identified in traditional macroeconomic models.

Inflation and the Dispersion of Prices. One well-known empirical regularity of inflation is that high rates of inflation are often associated with a greater dispersion of prices across goods in the economy; that is, high inflation causes the distribution of prices in the economy to widen.[7] This relationship is

[7] For evidence from historical hyperinflations, such as those experienced in post–World War I Germany, see Graham (1930) and Hercowitz (1981). Van Hoomissen (1988) provides empirical support for the more recent inflation in Israel.

evident not only in the distribution of relative prices of different goods, such as apples and oranges, but also in the distribution of the price of similar or identical goods across different sellers. For example, Wal-Mart and Kmart may charge different prices for a similar television set, and the evidence indicates that higher overall inflation will likely increase the difference in these prices. These effects of increasing uncertainty about the prices of goods and distorting relative prices are important welfare costs of inflation. Clearly, macroeconomic models of money that assume all individuals are identical and participate in perfectly competitive markets have little to say about how inflation affects price dispersion.[8]

A search model of money replicates a macroeconomic feature of money most economists would agree with: while short-run changes in money have real effects, long-run changes have primarily nominal effects.

Molico (1999) extends the prototype model with prices by allowing individuals to accumulate and store divisible units of currency. Since individuals are randomly meeting each other and accumulating and spending money at different times, there will be a distribution of money holdings across individuals. Buyers and sellers again negotiate prices in a way that divides the gains from trade between them, depending on how much money the buyer has available to trade. For example, if a buyer with ten dollars meets a seller, the benefit the buyer gets from spending an extra dollar to purchase a good will be different from that of someone with only two dollars. Consequently, prices will generally be different in each trade, depending on buyers' money holdings, and there will be price dispersion.

The model verifies that changes in the growth rate of the money supply, and hence inflation, can have important redistributive effects on money holdings across the population. In particular, Molico finds that if the new money is distributed to individuals in a lump sum, an increase in the rate of monetary expansion would decrease the dispersion of prices and improve welfare if inflation is sufficiently low. In contrast, increasing money growth in a high-inflation environment would increase price dispersion and hence lower welfare. Thus, injecting the economy with new cash has two opposing effects. First, it reduces income inequality by making individuals with low money holdings relatively better off than cash-rich individuals, thus reducing price dispersion. Second, inflation lowers the average amount of real money held per person proportionally to an individual's money holdings, thus increasing inequality and price dispersion. If the inflation rate is high enough, this latter effect can dominate the former.

These results conform with empirical studies from hyperinflation countries. They suggest that countries with very stable prices and low inflation, such as the United States, the United Kingdom, and Japan, may actually benefit from money growth because it will narrow price dispersion and wealth inequality. But in countries with high inflation, such as Germany after World War I and Latin American countries in the 1980s, inflation will exacerbate the dispersion of prices, with detrimental effects on economic activity and welfare.

Inflation and Capital Accumulation. Beginning with the works of Mundell (1963) and Tobin (1965), one of the central issues in analyzing inflation and economic activity is its impact on productive capital. The Mundell-Tobin effect asserts that inflation affects the portfolio decision, causing individuals to hedge by substituting out of cash and into productive assets. This effect suggests that inflation and capital investment may be positively related. Market-clearing models, such as cash-in-advance models, say that inflation taxes all activities requiring cash, including the purchase of capital goods. These models predict that the capital stock should fall with higher inflation.

Shi (1999) considers this question in the context of a search model in which money is divisible yet every household has identical money balances at every point in time. That is, there is no distribution of money holdings as in Molico's approach. Shi accomplishes this model by treating a household as consisting of many members, each of whom either is a seller or holds an indivisible amount of money. Thus, while money is indivisible to a member of the household, it is divisible to the household, which makes decisions about how many goods to purchase with money and how many goods to sell for money. Search and matching occurs among the members of different households.[9]

Money growth and inflation in this model have an impact on capital through both the quantity of goods exchanged in each meeting between a buyer and seller (the intensive margin) and the numbers of buyers and sellers participating in the market (the extensive margin). The intensive margin results tend to conform with the prediction of market-clearing models of money; inflation taxes money and consumption and makes leisure more attractive than labor. Since capital and labor complement each other in the production of goods and services, investment in productive capital declines as well. The extensive margin not present in these traditional models tell a different story, however. Just as in Li (1994, 1995), inflation creates an incentive to carry out transactions more quickly, increasing the number of buyers with money participating in the market. This rise increases the frequency of transactions and the incentives to produce and accumulate capital. If the inflation rate is low enough, the extensive effects of the model's search frictions can dominate the intensive effects. Thus, a Mundell-Tobin effect can exist for an appropriate range of low/moderate inflation rates, leading to a positive impact on productive activity.

Multiple Matching Models As an Alternative

Multiple matching models of money represent a new class of search-theoretic models of money designed with macroeconomic applications in mind. The framework, as developed by Laing, Li, and Wang (1999), departs from the prototype search models in some significant ways. Buyers are matched with a large (multiple) number of sellers instead of one to one. An analogy would be a shopper walking into a farmers market carrying goods and money and encountering many products and sellers. If the number of vendors in the farmers market is large enough, the buyer is always able to find barter and monetary trading opportunities. Since all individuals behave in a similar way, the advantage of our approach is that they will consume similar amounts and have identical money holdings (there are no distributions of money holdings or prices).[10]

Second, even though individuals still like to consume a small fraction of goods produced in the economy, they desire consumption variety and purchase a basket of goods. An example is a vegetarian who would prefer to buy carrots and broccoli instead of just broccoli. Because the number of sellers a buyer meets who satisfy the double coincidence of wants is small compared to those who may be willing to accept money, consumption variety is much smaller in a barter economy than in a monetary one. The use of money expands trading opportunities, allowing individuals to purchase a greater variety of goods.

Inflation, Employment, and Productivity Activity. Laing, Li, and Wang (2000) provide an example of how multiple matching models of money can shed new light on the links between monetary growth and productive activity. The study is motivated by empirical work showing that a consistently negative relationship between inflation, employment, and output is not found in the data across many countries. While it is true that periods of sustained inflation and hyperinflation disrupt productive activity, this observation tends not to apply to low-inflation countries (see Bullard and Keating 1995; Ahmed and Rogers 2000). To study this issue, Laing, Li, and Wang construct a multiple-matching model in which individuals must allocate their time between leisure, work, and investment in shopping (search) effort. Investing a greater amount of time in searching for goods, or shopping, increases the number of sellers one can meet in the farmers market and hence the variety of products available for purchase.

The results show that the importance of money in carrying out transactions plays a crucial role in how money growth and inflation affect productive activity. If money does not play an important role in overcoming trade frictions, or if these frictions are not too severe, a monetary expansion leads only to the inflation tax effect identified in traditional models of money. Individuals move away from market participation by both working and shopping less, and productive activity declines. However, if trade frictions are severe and money as a medium of exchange serves an important function, then increasing monetary growth can actually encourage productivity through its impact on trade and market participation. The intuition is that an increase in the inflation rate tends to erode purchasing power. One way individuals can compensate is to shop more intensely. By doing so, they meet with a greater number of sellers from whom to purchase greater consumption variety (that is, they compensate for the lower quantity of consumption with greater quality). This expansion of

[8] Search theory has been applied in a different way to explain this phenomenon. In such an approach, sellers may be able to charge different prices for the same good because they are in different locations and have local monopoly power (Fishman 1992; Benabou 1992). Individuals must expend shopping resources to "search" for the best price. However, these models really have no use for money and treat price inflation as exogenous; that is, they are not macroeconomic models of money.

[9] While the idea of a "large household" may at first seems strange and unrealistic, it appeals to the notion that the transactions between the members of a household in a given time period can be thought of as the many transactions an individual makes over the period as a buyer with money and a seller with goods.

[10] This result comes from the statistical "law of large numbers." While the idea is similar to Shi's (1997), it does not involve a fictitious large household, appealing to the more realistic notion that individuals engage in a large number of transactions over time.

trading opportunities increases individual and firm participation in both the goods and labor markets, and these increases translate to an increase in overall employment and output.[11] If inflation is not too high, this result is consistent with the empirical Phillips curve, which depicts a negative trade-off between inflation and unemployment. However, this result is driven purely by the explicit role of money as a medium of exchange, not by assumptions of wages and prices rigidities common in Keynesian-style models.

Conclusion

Search-theoretic models of money formalize the role of money as a medium of exchange often ignored or not rigorously modeled in the traditional macroeconomic literature. These models explicitly capture the trade frictions that motivate the use of money in transactions and offer new insights into classic issues in monetary economics and macroeconomics as well as some issues, such as the use of multiple currencies, that cannot be addressed by the traditional models.

For example, a number of search-theoretic models of money predict that while high inflation is undoubtedly disruptive to economic activity, a small but positive amount of money growth and inflation can have beneficial effects on output and welfare. These benefits arise from the effect money and inflation have on the liquidity of the transactions process. At a very intuitive level, this prediction conforms with a conventional wisdom that a small amount of inflation can "lubricate" the gears of economic activity through promoting exchange between individuals and firms. Macroeconomic models that simply assert a transaction role for money overlook this aspect of money and inflation. Furthermore, the prediction is consistent with recent empirical evidence from Ahmed and Rogers (2000), among others, suggesting that inflation has had a positive long-run impact on consumption, output, and investment in low-inflation countries like the United States.[12] The progress search-theoretic models of money have made so far in investigating these issues demonstrates that why people use money is important to macroeconomics.

These macroeconomic applications also reveal the enormous potential of search models of money to address an even broader range of issues. So far, most of the results of these applications are qualitative. A quantitative analysis may not only explore the issue of how well these models explain the empirical facts regarding money and economic activity but may also provide guidelines for the operation of monetary policy.

For example, a natural question to study is, What is the "optimal rate of inflation" policymakers should target? Another related issue is, What do these models have to say about the interrelationship between money and credit markets? Thus far, work in this area has focused primarily on the nature of credit arrangements in search economies (Diamond 1990; Hendry 1992; Corbae and Ritter 1997), conditions under which individuals use both monetary and credit exchange simultaneously (Aiyagari, Wallace, and Wright 1996), and the role for privately issued currency, an issue that was especially important before the establishment of central banking in the United States (Cavalcanti, Erosa, and Temzelides 1999; Monnet 2001). These works suggest that search-theoretic models should be well suited to studying the economic consequences of the evolution of the payments system—for example, fiat money versus checks versus electronic money—and the role credit markets play in the way monetary policy affects macroeconomic activity.

[11] These findings also indicate that in some situations, the model predicts that money growth may have either a positive or negative impact on productive activity, or multiple equilibria. This result may also explain why empirical work has failed to identify a consistent relationship between money growth and economic activity in low-inflation countries.

[12] Ahmed and Rogers (2000) use over one hundred years of U.S. data to analyze the long-run empirical relationship between inflation and economic activity.

REFERENCES

AHMED, SHAGHIL, AND JOHN H. ROGERS. 2000. Inflation and the great ratios: Long-term evidence from the United States. *Journal of Monetary Economics* 45 (February): 3–35.

AIYAGARI, RAO, NEIL WALLACE, AND RANDALL WRIGHT. 1996. Coexistence of money and interest bearing securities. *Journal of Monetary Economics* 37 (June): 397–419.

BALASSA, BELA. 1964. The purchasing power parity doctrine: A reappraisal. *Journal of Political Economy* 72 (December): 584–96.

BARRO, ROBERT J. 1976. Rational expectations and the role of monetary policy. *Journal of Monetary Economics* 1 (January): 1–32.

BENABOU, ROLAND. 1992. Search, price setting, and inflation. *Review of Economic Studies* 55 (July): 353–76.

BULLARD JAMES, AND JOHN KEATING. 1995. The long-term relationship between inflation and output in postwar economies. *Journal of Monetary Economics* 36 (December): 477–96.

CAMERA, GABRIEL, AND DEAN CORBAE. 1999. Money and price dispersion. *International Economic Review* 40 (November): 985–1008.

CAVALCANTI, RICARDO DE O., ANDRES EROSA, AND TED TEMZELIDES. 1999. Private money and reserve management in a random matching model. *Journal of Political Economy* 107 (October): 929–45.

CORBAE, DEAN, AND JOSEPH RITTER. 1997. Money and search with enduring relationships. University of Pittsburgh. Unpublished paper.

CRAIG, BEN, AND CHRISTOPHER WALLER. 1999. Dual-currency economies as multiple-payment systems. Federal Reserve Bank of Cleveland *Economic Review* (First Quarter): 2–13.

CURTIS, ELIZABETH SOLLER, AND CHRISTOPHER WALLER. 2000. A search-theoretic model of legal and illegal currency. *Journal of Monetary Economics* 45 (February): 155–84.

DIAMOND, PETER. 1990. Pairwise credit in search equilibrium. *Quarterly Journal of Economics* 105 (May): 285–319.

FISHMAN, ARTHUR. 1992. Search technology, staggered price-setting, and price dispersion. *American Economic Review* 82 (March): 287–98.

FRIEDMAN, MILTON. 1968. The role of monetary policy. *American Economic Review* 58 (March): 1–17.

GRAHAM, FRANK D. 1930. *Exchange, prices, and production in hyper-inflation: Germany 1920–1923.* Princeton, N.J.: Princeton University Press.

HENDRY, SCOTT. 1992. Credit in a search model with money as a medium of exchange. University of Western Ontario. Unpublished paper.

HERCOWITZ, ZVI. 1981. Money and the dispersion of relative prices. *Journal of Political Economy* 89 (April): 328–56.

JEVONS, WILLIAM STANLEY. 1875. *Money and the mechanism of exchange*. London: Appleton.

KIYOTAKI, NOBUHIRO, AND RANDALL WRIGHT. 1989. On money as a medium of exchange. *Journal of Political Economy* 97 (August): 927–54.

———. 1991. A contribution to the pure theory of money. *Journal of Economic Theory* 53 (April): 215–35.

———. 1993. A search-theoretic approach to monetary economics. *American Economic Review* 83 (March): 63–77.

LAING, DEREK, VICTOR LI, AND PING WANG. 1999. Money and prices in a multiple matching model of money. Pennsylvania State University. Unpublished paper.

———. 2000. Inflation and productive activity in a multiple matching model of money. Princeton University. Unpublished paper.

LI, VICTOR E. 1994. Inventory accumulation in a search-based monetary economy. *Journal of Monetary Economics* 34 (December): 511–36.

———. 1995. The optimal taxation of fiat money in search equilibrium. *International Economic Review* 36 (November): 927–42.

LUCAS, ROBERT E. 1973. Some international evidence on output-inflation trade-offs. *American Economic Review* 63 (June): 326–34.

MATSUYAMA, KIMINORI, NOBUHIRO KIYOTAKI, AND AKIHIRO MATSUI. 1993. Toward a theory of international currency. *Review of Economic Studies* 60 (April): 283–307.

MOLICO, MIGUEL. 1999. The distribution of money and prices in search equilibrium. University of Western Ontario. Unpublished paper.

MONNET, CYRIL. 2001. Optimal public money. University of Minnesota. Unpublished paper.

MUNDELL, ROBERT. 1963. Inflation and real interest. *Journal of Political Economy* 71:280–83.

ROGOFF, KENNETH. 1996. The purchasing power parity puzzle. *Journal of Economic Literature* 34 (June): 647–68.

SHI, SHOUYONG. 1995. Money and prices: A model of search and bargaining. *Journal of Economic Theory* 67 (December): 467–96.

———. 1997. A divisible search model of fiat money. *Econometrica* 65 (January): 75–102.

———. 1999. Search, inflation, and capital accumulation. *Journal of Monetary Economics* 44 (August): 81–103.

TOBIN, JAMES. 1965. Money and economic growth. *Econometrica* 32 (October): 241–47.

TREJOS, ALBERTO, AND RANDALL WRIGHT. 1993. Search, bargaining, money, and prices: Recent results and policy implications. *Journal of Money, Credit, and Banking* 25 (August): 558–76.

———. 1995. Search, bargaining, money, and prices. *Journal of Political Economy* 103 (February): 118–41.

———. 1996. Search-theoretic models of international currency. Federal Reserve Bank of St. Louis *Review* 78 (May): 117–32.

———. 2000. International currency. University of Pennsylvania. Unpublished paper.

VAN HOOMISSEN, THERESA. 1988. Price dispersion and inflation: Evidence from Israel. *Journal of Political Economy* 96 (December): 1303–14.

WALLACE, NEIL. 1997. Short-run and long-run effects of changes in money in a random matching model. *Journal of Political Economy* 105 (December): 1293–1307.

WICKSELL, KNUT. 1911. *Lectures on Political Economy: Money*. Vol. 2. New York: Kelley.

ZHOU, RUILIN. 1997. Currency exchange in a random search model. *Review of Economic Studies* 64 (April): 289–310.

The Risks and Rewards of Selling Volatility

SAIKAT NANDI AND DANIEL WAGGONER

Nandi is a former senior economist at the Atlanta Fed and is currently a financial engineer at Fannie Mae. Waggoner is an economist in the Atlanta Fed's research department. They thank Jerry Dwyer and Larry Wall for helpful comments.

BUYING AND SELLING CERTAIN KINDS OF VOLATILITY-SENSITIVE OPTIONS PORTFOLIOS IS A POPULAR PRACTICE EVEN THOUGH THIS ACTIVITY IS ASSOCIATED WITH SUBSTANTIAL RISK. WHEN SUCH PORTFOLIOS ARE FORMED, TWO FACTORS SIGNIFICANTLY INFLUENCE THE VALUE OF THE OPTIONS: THE PRICE OF THE UNDERLYING ASSET AND THE FUTURE VOLATILITY EXPECTED TO prevail until the options expire. It is possible to form a portfolio of call and put options so that the portfolio's payoff is very sensitive to the volatility of the underlying asset but only minimally sensitive to changes in the level of the underlying asset. Traders and investors who frequently buy or sell such portfolios do so with a view of the volatility of the underlying asset that does not correspond to the expected future volatility embedded in the option price, or the implied volatility.[1] For example, the former hedge fund Long Term Capital Management had created portfolios of options on certain stock indexes based on its view that expected future volatility was different from the prevailing implied volatilities (Dunbar 1999).

There is often substantial risk, however, in options portfolios set up to exploit a perceived mispricing in the expected volatility of the underlying asset without initial sensitivity to the level of the asset. This risk stems from possible subsequent abrupt changes in the level of the asset price. Changes in volatilities are highly negatively correlated with changes in levels in many asset markets. If a short position in implied volatility on a market is created and a subsequent sharp market decline results in much higher implied volatility, the value of the implicit short position in volatility could dramatically decrease. For example, the former Barings PLC sustained huge losses from short positions in volatility (initiated by a trader, Nick Leeson) on Nikkei futures as the Nikkei plunged in early 1995 (Jorion 2000).

The objective of this article is to delineate the risks and rewards associated with the popular practice of selling volatility through selling a particular portfolio of options called straddles. Toward this end, the article first examines the statistical properties of the returns generated by selling straddles on the Standard and Poor's (S&P) 500 index. Although it is theoretically possible to construct other options portfolios to make volatility bets (Carr and Madan 1998), straddles are by far the most popular type of portfolio. The article also demonstrates that the usual practice of selling volatility by comparing the observed implied volatility (from option prices) with the volatility expected to prevail (given the history of asset prices) could be flawed. This flaw could arise if the underlying asset has a positive risk premium—that is, if its expected return over a given horizon exceeds the risk-free rate over the same horizon—and the returns of the underlying asset are negatively correlated with changes in volatility. Thus, basing the decision to sell a straddle on a comparison of seemingly irrational high implied volatilities with much lower expected volatility could itself be an irrational choice.

Straddles

Straddles are very popular but quite volatility-sensitive options portfolios. A straddle consists of a call and put option of the same strike price and maturity so that the strike price is equal to (or very close to) the current price of the underlying asset. The higher the volatility until the option expires, the higher the expected profits from buying the straddle, and vice versa. For example, suppose an investor buys a European straddle that matures in fifty days, the current price of the underlying stock is $100, and the strike prices of both the European call and put options are $100. Under the (BSM) Black-Scholes-Merton model (outlined in Black and Scholes 1973 and Merton 1973) with an annualized implied volatility of 20 percent and a risk-free rate of 5 percent, the price of the call option is $3.29, and the put option price is $2.61. The cost of the straddle is thus $5.90 ($3.29 + $2.61). If the stock price at the expiration of the option is at least $105.90 (payoff from the call option is $5.90, and payoff from the put option is zero) or $94.10 (payoff from the put option is $5.90, and payoff from the call option is zero), the buyer breaks even. The greater the stock price is than $105.90 or the lower it is then $94.10, the greater the buyer's profits are. Conversely, if the stock price is greater than $94.10 but less than $105.90, the seller of the straddle turns in a profit because the revenue generated by the sales exceeds the losses when the straddle expires. Of course, the narrower the dispersion of the possible stock prices at maturity, the higher the seller's profits. The chart shows the payoff from buying the straddle when the options expire.

A straddle might seem to be an ideal vehicle through which to express a view on the future volatility of the underlying asset without necessarily having an initial view on the future direction of the asset price.

When the straddle is created, the change in its value is not very sensitive to the change in the value of the underlying asset. The previous example and the chart clearly show, however, that the higher the volatility of the underlying asset between the present time and option expiration, the higher the potential payoff from a long position in the straddle. A straddle might thus seem to be an ideal vehicle through which to express a view on the future volatility of the underlying asset without necessarily having an initial view on the future direction of the asset price—that is, a trader might buy a straddle if volatility is expected to be high or sell one if volatility is expected to be low.

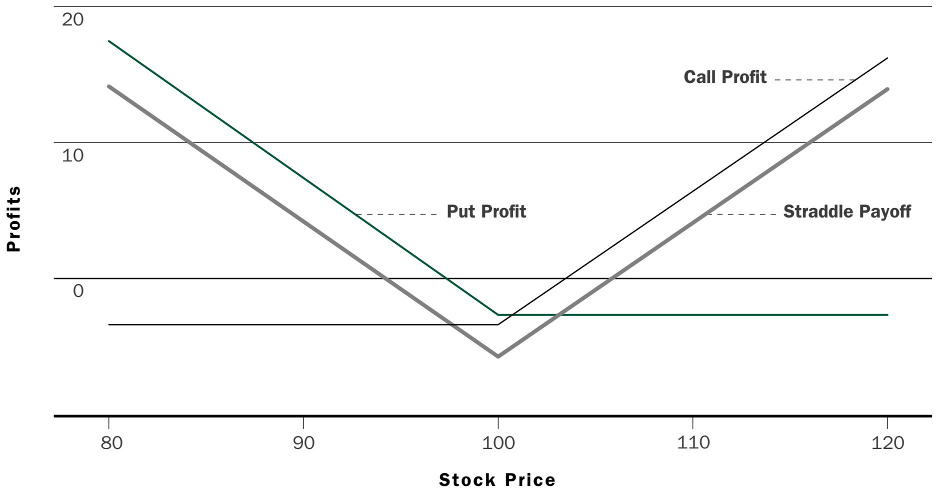
To see how a straddle is a bet on volatility, assume that the implied volatility (using the BSM model) from a straddle on an equity option that expires in one month is 40 percent (annualized). On the other hand, assume that the maximum historical volatility of the underlying asset that has been observed over any one-month horizon is 30 percent (annualized). Based on this observation, would one sell the straddle—in other words, sell volatility—because it appears to be high? If volatility is constant or evolves deterministically, that is, if the assumptions of the BSM model hold true, it may be tempting to sell the straddle. It has been well documented, however, that in most asset markets volatility evolves randomly through time (Bollerslev, Chou, and Kroner 1992). Even if volatility were constant or predictable, the price of the underlying asset could periodically undergo an abrupt level shift—sometimes referred to as a jump in the price process.[2] Because of these factors, a risk premium related to the randomness of volatility or unpredictable jumps in asset prices could get incorporated into the price of an option (Bates 1996). The implied volatility determined from the price of an option using the BSM model would thus be contaminated with the relevant risk premium, making any comparison with historical volatility problematic.[3]

Box 1 gives an example in which the volatility of the underlying asset evolves randomly and the changes in volatility and returns are negatively correlated. The box shows that as long as the underlying asset has a positive risk premium, the implied volatility from the options could be higher than a historical measure of volatility. Because the BSM model may be an inadequate description of the observed option prices, caution is warranted in using it to compare implied volatility to historical volatility.

The Implied and Historical Volatility of the S&P 500

In trading straddles on a market index such as the S&P 500, it is important to have a good understanding of the dynamics of volatility through time. This section offers a computation of the implied

A Hypothetical Payoff Scenario



Note: Chart shows profits at option expiration from buying the straddle as well as the profits at expiration from buying a call and put option with strike prices of 100.

volatilities from near-the-money options (options in which the strike price is as close as possible to the price of the underlying asset) on the S&P 500 and a simple but frequently used measure of historical volatility for each trading day over a six-year horizon, 1990 through 1995.[4] On each day, historical volatility is computed as the standard deviation from a sample of the last thirty days of daily S&P 500 returns (close-to-close) and then multiplied by the square root of 252 (as there are about 252 trading days in a year with daily returns assumed to be independently and identically distributed).

The means and standard deviations of the annualized implied volatilities and the annualized historical volatilities are shown in Table 1. This table demonstrates that the implied volatility from near-the-money S&P 500 index options tends to be above (on average, 3 percent) the measure of historical volatility frequently used in practice.[5] In addition, the minimum and maximum historical volatilities observed over this period tend to be lower than minimum and maximum implied volatilities. Someone comparing the implied volatility with the simple measure of historical volatility might often be tempted to sell straddles. Box 1 shows, however, that implied volatilities can be above historical volatilities without any trading opportunities.

The Impact of Correlation and Implied Volatility Skew

Although it is possible to sell straddles based on an incorrect comparison of volatilities, the larger risk from selling straddles often arises from the correlation between returns and volatility as well as large changes in asset price over a short period. In many equity markets, returns and volatility tend to be negatively correlated—volatility goes up if returns go down, and vice versa. From 1990 to 1995, the average correlation between daily returns on the S&P 500 and the daily changes in implied volatilities was –0.525. For example, on August 23, 1990, as the S&P 500 closed at 307.06, down from its previous

[1] Because a very high implied volatility (currently observed from a near-the-money option) is likely to revert to a much lower level over a period of time—a phenomenon known as mean reversion in volatility—selling option portfolios to take advantage of this mean reversion may seem appealing.

[2] The stock market crash of 1987 could be viewed as a jump in the price process of an asset.

[3] Even if the payoff from the option can be perfectly replicated by trading in the underlying asset and the risk-free asset, as in the BSM model (that is, the option is a redundant security), a risk premium that is related to the riskiness of the underlying asset would still show up in the option values and therefore in the implied volatilities as long as volatility is path-dependent (Heston and Nandi 2000).

[4] Options on the S&P 500 are traded at the Chicago Board Options Exchange. The description of the data from which these implied volatilities are computed appears in Box 2.

[5] However, if a measure of historical volatility were computed using a model that permits the volatility to be time-varying and unpredictable, it would not necessarily lie below the implied volatility.

BOX 1

Comparing Historical Implied Volatility

If volatility is random or path-dependent, and asset returns and volatilities are negatively correlated, then implied volatilities from option prices could be higher than the observed historical volatility from the underlying asset. Assume that the natural logarithm of the asset price conforms to the following GARCH-type process (Heston and Nandi 2000) over time steps of length Δ:

$$\log[S(t)] = \log[S(t-\Delta)] + r + \lambda h(t) + \sqrt{h(t)}\,z(t),$$

$$h(t) = \omega + \beta h(t-\Delta) + \alpha[z(t-\Delta) - \gamma\sqrt{h(t-\Delta)}]^2,$$

where r is the continuously compounded risk-free rate for the time interval Δ, $z(t)$ is a standard normal disturbance, and $h(t)$ is the conditional variance of the log return between $t - \Delta$ and t and is known from the information set at time $t - \Delta$. In this model, if γ is positive, there is negative correlation between asset returns and volatility. In particular, the variance, $h(t)$, depends on the entire path of asset prices until time $t - \Delta$.

Options are typically valued not under the distribution that generates the observed asset prices but under an adjusted distribution called the risk-neutral distribution (Cox and Ross 1976; Harrison and Kreps 1979). Heston and Nandi (2000) show that under negative correlation between returns and volatility ($\gamma > 0$), as long as the expected return of the risky asset exceeds the risk-free rate, the conditional mean of the variance, $h(t)$, under the risk-neutral distribution exceeds the conditional mean of $h(t)$ under the data-generating distribution. The drift of the variance under the risk-neutral distribution tends to be higher than the drift of the variance under the data-generating distribution. Since implied volatility (to a large extent) reflects the drift or the expected value of the variance under the risk-neutral distribution, it is quite possible for the implied volatility to be higher than the expected volatility.[1]

[1] This assertion ignores the small bias that can arise from Jensen's inequality as option values are nonlinear functions of volatility. However, the degree of nonlinearity tends to be small for at-the-money options.

overnight close of 316.55, the implied volatility from short-maturity straddles (seven to forty days in expiration) went up to 0.322 from the previous day's 0.265.[6] Similarly, on August 27, 1990, as the S&P 500 closed at 321.44, up from its overnight close of 311.31, the implied volatility from the straddles went down to 0.239 from previous day's 0.307.

If a trader sells volatility through straddles and the market goes down, the value of the short position in the straddles drops considerably, more than the loss incurred on a long position in the market. For example, if a trader had sold a straddle on the S&P 500 index at the beginning of the day on November 11, 1991, and held it through the day as the market went down around 3.6 percent, the action would have incurred a loss of around 9 percent on the straddle position over the day. At the same time, sharp upswings in the asset price could also adversely impact the value of a straddle. As the S&P 500 increased from 343.95 to 356.95 from January 31, 1991, to February 7, 1991, the return from selling straddles on January 31 and closing out on February 7, was –12.2 percent.

Another feature of implied volatilities that could adversely impact the seller of the straddle if the price of the underlying asset goes up is the so-called skew or smirk in implied volatilities. It is well known that the S&P 500 index options market consistently exhibits a skew in implied volatilities—the lower the strike price of an option, the higher the implied volatility, and vice versa (Rubinstein 1994; Nandi and Waggoner 2000). If the S&P 500 rises sharply, the near-the-money call option of the straddle becomes an in-the-money call (the strike is less than the price of the underlying asset). The loss in the value of the short-call position thus would stem not only from the change in the level of the S&P 500 but also perhaps from the skew in implied volatilities because an in-the-money call has a higher implied volatility than a near-the-money call.

Selling Straddles on the S&P 500 Index

An empirical exercise helps gauge the actual risks and rewards associated with selling straddles in a liquid options market over a long enough period. In this empirical exercise, a

TABLE 1
Statistics Computed from Near-the-Money S&P 500 Index Options, 1990–95

	Mean	Standard Deviation	Min	Max
Implied Volatility	0.136	0.039	0.074	0.325
Historical Volatility	0.106	0.036	0.050	0.226

TABLE 2 Returns from Unconditionally Selling Straddles

Days	Annual Mean	Annual Standard Deviation	Skewness	Max	Min
			Short-Maturity Straddles		
10	0.231	0.228	–1.79	0.117	–0.292
15	0.385	0.239	–1.67	0.162	–0.295
20	0.334	0.289	–2.12	0.136	–0.424
			Medium-Maturity Straddles		
10	0.106	0.187	–2.25	0.116	–0.282
15	0.225	0.221	–2.49	0.145	–0.373
20	0.279	0.239	–2.23	0.172	–0.416

trader sells straddles of two different maturities in the market for S&P 500 index options and then liquidates the positions (buys back the straddles) after a certain number of days.[7] The market for S&P 500 options is one of the most active options markets in the world. (The description of options data used for the empirical exercise for the 1990–95 period appears in Box 2.) This exercise of selling straddles adheres to all the margin requirements (initial margin and maintenance margin) of the Chicago Board Options Exchange (where S&P 500 options are traded). The round-trip transactions cost in the form of bid-ask spreads is also explicitly recognized because these spreads are not insignificant.

As a first exercise, the trader sells the straddles every day (1990–95) without comparing their implied and historical volatilities. Table 2 shows some statistics of holding period returns generated by selling short- and medium-maturity straddles at the bid prices and then liquidating the positions (buying back the options at the ask prices) after ten, fifteen, and twenty trading days.[8] In computing these returns, the initial margin the trader must put up according to exchange rules is treated as the initial investment. The mean and standard deviations of the holding period returns shown in these tables are annualized. The average annual returns from short- and medium-maturity straddles that are liquidated after twenty days are 33.4 percent and 27.9 percent with annualized standard deviations of 28.9 percent and 23.9 percent. In contrast, the corresponding average annual return from holding the S&P 500 index for twenty trading days over this period is 12.9 percent with an annualized standard deviation of 10.9 percent.

This exercise shows that although selling the straddle may appear to achieve risk-adjusted returns comparable to holding the market if standard deviation is used as the measure of risk, the returns from selling the straddles are much more negatively skewed than returns from holding the market: the coefficients of skewness from selling the short- and medium-maturity straddles are –2.12 and –2.23 while that from holding the market is 0.075. The probability of negative returns exceeds the probability of positive returns of equal magnitude. Further, the probability of large negative returns far exceeds

[6] The strike price of these straddles is set so that the options are at-the-money-forward, that is, $K = S(t)\exp[r(\tau)\tau]$, where K is the strike price, $S(t)$ is the current asset price, τ is the time to maturity of the option, and $r(\tau)$ is the risk-free interest rate applicable for the maturity.

[7] The straddle sold resembles as closely as possible an at-the-money-forward straddle.

[8] Short-maturity straddles have maturities of twenty to forty calendar days, and medium-maturity straddles have maturities of forty to seventy-five calendar days.

BOX 2

S&P 500 Data Used in Creating Straddles

The market for S&P 500 index options is one of the most active index options market in the United States; it is also the largest in terms of open interest in options. The data set used for creating the straddles is a subset of the tick-by-tick data on the S&P 500 options that includes both the bid-ask quotes and the transaction prices; the raw data set was obtained directly from the exchange.

As many of the stocks in the S&P 500 index pay dividends, a time series of dividends for the index is required for computing straddles. This case uses the daily cash dividends for the S&P 500 index collected from the S&P 500 information bulletin. The present value of the dividends is subtracted from the current index level. For the risk-free rate, the continuously compounded Treasury bill rate (from the average of the bid and ask discounts reported in the *Wall Street Journal*), interpolated to match the maturity of the option, is used. Because the S&P 500 level that appears with the options records may be stale, this article's computations use index levels implied from the S&P 500 futures traded at the Chicago Mercantile Exchange. The nearest matches for maturity of S&P 500 futures prices (in terms of the difference in time stamps between the options record and the futures record) are used to get the implied S&P 500 index levels. These futures prices are created from tick-by-tick S&P 500 futures data sets obtained from the Futures Industry Institute. Given a discrete dividend series, the following equation (Hull 1997) is used to compute the implied spot price (S&P 500 index level): $F(t) = [S(t) - PVDIV]e^{r(t)(T-t)}$, where $F(t)$ denotes the futures price, $PVDIV$ denotes the present value of dividends to be paid from time t until the maturity of the futures contract at time T, and $r(t)$ is the continuously compounded Treasury bill rate (from the average of the bid and ask discounts reported in the *Wall Street Journal*), interpolated to match the maturity of the futures contract. In terms of maturity, options with less than six days or more than one hundred days to maturity are excluded.[1]

An option of a particular moneyness and maturity is represented only once in the sample on any particular day. Although the same option may be quoted again in this time window (with same or different index levels) on a given day, only the first record of that option is included in this sample for that day.

A transaction must satisfy the no-arbitrage relationship (Merton 1973) in that the call price must be greater than or equal to the spot price minus the present value of the remaining dividends and the discounted strike price. Similarly, the put price has to be greater than or equal to the present value of the remaining dividends plus the discounted strike price minus the spot price.

The call and put options that the straddle comprises are chosen each day from the bid-ask quotes after 9:00 A.M. central standard time (CST) so that they are closest to being at-the-money-forward; that is, $K = S(t)\exp[r(\tau)\tau]$, where K is the strike price, $S(t)$ is the current asset price, τ is the time to maturity of the option, and $r(\tau)$ is the risk-free interest rate applicable for the maturity. Only one straddle is formed each day, and on the day of the liquidation (when the straddle is bought back), the option positions are closed out from the bid-ask quotes after 9:00 A.M. CST.

[1] See Dumas, Fleming, and Whaley (1998) for a justification of the exclusionary criteria about moneyness and maturity.

the probability of large positive returns. For example, the highest twenty-day return (not annualized) from selling short maturity straddles is 13.6 percent whereas the lowest twenty-day return is –42.4 percent. How important is the negative correlation between returns and volatility in determining some of the extreme negative returns? When the S&P 500 lost around 8 percent over a fifteen-day trading period starting in July 20, 1990, the implied volatility spiked from 14 percent to around 25.5 percent (a change of around 82 percent), and the return from selling a straddle was –28 percent.

The returns from short positions in the straddles could result from a change in the level of the market or changes in the implied volatilities of the options constituting the straddles. The high returns from selling straddles are related to the correlations between returns from short positions in straddles and the returns on the S&P 500, as well as the correlations between straddle returns and changes in the implied volatilities over the holding period of the options. In Table 3, $\rho(R_s, R_m)$ denotes the correlation between returns from selling straddles and returns from buying the S&P 500 over the holding period, and $\rho(R_s, IV)$ denotes the correlation between returns from selling the straddles and the changes in the implied volatilities of the options. Table 3 also shows the correlation between returns from selling medium-maturity straddles and buying the S&P 500 as well as the correlation between returns from selling medium-maturity straddles and changes in implied volatilities.

Short positions in straddles stand to gain mainly from the decrease in implied volatilities over the holding period. The correlations between the returns from selling the straddles and market returns tend to be negative, but for some holding periods they are only slightly negative.[9] In contrast, the correlations between straddle returns and changes in implied volatilities are substantially negative across all holding periods. This finding suggests that the positive returns from the short straddle positions mainly result from decreases in implied volatilities of the options in the straddle. Similarly, the negative returns from selling the straddles primarily result from the increases in implied volatilities of the options in the straddle.

Often, straddles are sold only if the implied volatility from the options exceeds the historical volatility that has been observed over a certain period of time. Thus, instead of selling straddles unconditionally, it is less risky to sell the straddle only if both the put and call implied volatility exceed the thirty-day historical volatility by at least 1 percent.[10] Table 4 shows some of the returns generated by selling short-maturity straddles using this decision rule, based on comparisons between implied and historical volatilities.

The mean returns are actually lower using the trading rule than they would be if the straddles were sold unconditionally. For example, if we liquidate the straddles after fifteen trading days, the mean (annualized) return under the trading rule is 26.1 percent compared to 38.5 percent without the trading rule. At the same time, however, these data suggest that using the mechanical trading rule to sell straddles is no better than selling the straddles unconditionally: the standard deviations of the returns are a little higher whereas the coefficient of skewness (of the returns) is a little lower using the trading rule. Further, substantial downside risk still remains—the minimum return from selling implied volatility on the S&P 500 over a fifteen-day holding period is –29.5 percent, much lower than could be achieved by being either long or short on the S&P 500 over the same holding period in our sample.

TABLE 3
Correlation between Returns

Days	$\rho(R_s, R_m)$	$\rho(R_s, IV)$
Short-Maturity Straddles		
10	–0.05	–0.61
15	–0.28	–0.71
20	–0.38	–0.89
Medium-Maturity Straddles		
10	–0.09	–0.74
15	–0.08	–0.73
20	–0.20	–0.72

Does Rebalancing Help?

In the preceding trading exercises, as the level of the S&P 500 varied from the strike prices of the options constituting the straddle, the trader did not rebalance the straddle to make it delta-neutral. Delta measures the change in the value of the option relative to the change in the value of the underlying asset (see Box 3 for the delta of a straddle).[11] The straddle is only approximately delta-neutral at the beginning, and the value of the straddle becomes sensitive to changes in the level of the S&P 500 as the index starts moving away from its initial level. It is possible that some of the excessive skewness and some of the extreme negative returns from selling the straddles can be reduced or eliminated by rebalancing the straddle to be delta-neutral each day. An option valuation model (such as the BSM model) should be used to arrive at the number of units of the underlying asset that need to be bought or sold

[9] One should not necessarily conclude that the correlations between returns from buying the straddles and buying the S&P 500 are positive because initial margins and maintenance margins are taken into account in computing the returns from selling straddles. Buying a straddle does not require any initial or maintenance margins.

[10] Historical volatility was also computed from the last forty-five days and sixty days of returns, but the results were not significantly different. Similarly, more sophisticated volatility models, such as the asymmetric GARCH models used in Engle and Ng (1993) and Heston and Nandi (2000), have been used to construct various measures of volatility without any substantial differences in results.

[11] Although a straddle may have a low delta, it has a high gamma (that is, the rate of change in delta is high), especially at short maturities.

TABLE 4
Returns from Conditionally Selling Short-Maturity Straddles

Days	Annual Mean	Annual Standard Deviation	Skewness	Max	Min
10	0.181	0.222	–1.24	0.103	–0.181
15	0.261	0.271	–1.33	0.161	–0.295
20	0.199	0.319	–2.06	0.129	–0.413

TABLE 5
Returns from Selling Short-Maturity Straddles with Rebalancing

Days	Annual Mean	Annual Standard Deviation	Skewness	Max	Min
10	0.184	0.364	–1.45	0.265	–0.415
15	0.352	0.542	–0.93	0.638	–0.971
20	0.379	0.611	–0.45	0.854	–0.732

to rebalance the straddle. If the theoretical model used for rebalancing is different from the model generating the observed option prices, however, a delta-neutral position will not really result from rebalancing. Thus, the value of the portfolio may be adversely affected both by a change in the level of the market and volatility. In short, selling straddles and rebalancing to maintain delta-neutrality could expose the seller to model risk.

In the final exercise, in addition to selling a straddle, a certain number of units of the S&P 500 are bought or sold every day according to the BSM model so that the straddle stays delta-neutral.[12] Table 5 shows the various statistics of the three different holding-period returns of the short-maturity straddles from this exercise. (Note that the initial and final values of the options, the initial margin requirement, and cash flows that would accrue from buying/selling the S&P 500 are taken into account in computing the returns.) The table shows that rebalancing reduces the negative skewness of the returns from selling the straddles but makes the standard deviation of the returns go up. Thus, after taking into account both the standard deviation and skewness of returns, rebalancing the portfolio to be delta-neutral does not substantially alter the risk profile from selling straddles.

Since the number of units of the underlying asset that need to be bought or sold for rebalancing is derived from a theoretical option valuation model, the process of rebalancing is subject to model risk. Choosing a valuation model that does not capture the dynamics of option prices very well can lead to results that are no better or worse than without rebalancing. In this case, the BSM model is used. However, existing research (Bakshi, Cao, and Chen 1997; Nandi 1998; Dumas, Fleming, and Whaley 1998) suggests that even more sophisticated rebalancing schemes such as those using stochastic volatility models may not lead to substantially better results than the simple BSM model in the S&P 500 index options market.

Conclusion

Selling market volatility through selling straddles exposes traders and investors to substantial risk, especially in equity markets. Selling straddles has resulted in substantial losses at banks and hedge funds such as the former Barings PLC and Long Term Capital Management. Although the returns from selling straddles can sometimes be very lucrative, especially if the volatility at which the options are sold quickly reverts to a much lower level, the probability of large negative returns far exceeds the probability of large positive returns. Moreover, the negative correlation between equity market returns and implied volatilities could make the straddle values highly sensitive to the direction of the market. In other words, if a trader sells the volatility implicit in option prices and the market subsequently goes down, the mark-to-market value of the portfolio could significantly decrease as the implied volatility increases. While rebalancing the straddle to maintain minimal exposure to the direction of the market is theoretically feasible, the rebalancing process exposes a trader to model risk and may not always help. In short, selling volatility through selling straddles can be lucrative but is quite risky.

[12] See Nandi and Waggoner (2000) for an example that shows how to make a portfolio of options delta-neutral.

BOX 3

The Delta of a Straddle

A short-maturity, at-the-money-forward straddle has a low delta under the BSM model. If Δ_c is the delta of the call option and Δ_p is the delta of the put option, then the delta of the straddle, Δ_s, is

$$\Delta_s = \Delta_c + \Delta_p = 2\,N(d1) - 1,$$

where $N(d1)$ is the standard normal distribution function and $d1 = 0.5\sigma\sqrt{\tau}$, with σ being the volatility and τ being the time to expiration of the option. Expanding $N(d1)$ using a first-order Taylor expansion around zero yields

$$\begin{aligned} N(d1) &= N(0) + (0.5\sigma\sqrt{\tau})/\sqrt{2\pi} \\ &= 0.5[1 + (\sigma\sqrt{\tau})/\sqrt{2\pi}]. \end{aligned}$$

Hence, $\Delta_s \approx (\sigma\sqrt{\tau})/\sqrt{2\pi}$.

As an example, the table shows the delta of a straddle with thirty days ($\tau = 30/365$) to maturity for a few values of annualized σ.

Delta of a Straddle with Thirty Days to Maturity

Value of Annualized Volatility (σ)	Delta (Δ_s)
0.10	0.0114
0.15	0.0171
0.20	0.0228

REFERENCES

BAKSHI, GURDIP, CHARLES CAO, AND ZHIWU CHEN. 1997. Empirical performance of alternative option pricing models. *Journal of Finance* 52 (December): 2003–49.

BATES, DAVID. 1996. Jumps and stochastic volatility: Exchange rate processes implicit in deutsche mark options. *Review of Financial Studies* 9 (Spring): 69–107.

BLACK, FISCHER, AND MYRON SCHOLES. 1973. The pricing of options and corporate liabilities. *Journal of Political Economy* 81 (May/June): 637–54.

BOLLERSLEV, TIM, RAY CHOU, AND KENNETH KRONER. 1992. ARCH modeling in finance. *Journal of Econometrics* 52 (April/May): 5–59.

CARR, PETER, AND DILIP MADAN. 1998. Towards a theory of volatility trading. In *Volatility: New estimation techniques for pricing derivatives*, edited by Robert Jarrow. London: Risk Publications.

COX, JOHN, AND STEPHEN ROSS. 1976. The valuation of options for alternative stochastic processes. *Journal of Financial Economics* 3 (January/March): 145–66.

DUMAS, BERNARD, JEFF FLEMING, AND ROBERT WHALEY. 1998. Implied volatility functions: Empirical tests. *Journal of Finance* 53, no. 6:2059–2106.

DUNBAR, NICHOLAS. 1999. LTCM and the dangers of marking to market. *Risk* 12:6–7.

ENGLE, ROBERT, AND VICTOR NG. 1993. Measuring and testing the impact of news on volatility. *Journal of Finance* 43 (December): 1749–78.

HARRISON, J.M., AND D. KREPS. 1979. Martingales and arbitrage in multiperiod securities markets. *Journal of Economic Theory* 20 (June): 381–408.

HESTON, STEVEN L., AND SAIKAT NANDI. 2000. A closed-form solution GARCH option valuation model. *Review of Financial Studies* 13 (Fall): 588–625.

HULL, JOHN. 1997. *Options, futures, and other derivatives.* New Jersey: Prentice-Hall, Inc.

JORION, PHILIPPE. 2000. *Value at risk: The new benchmark for managing financial risk*. New York: McGraw Hill.

MERTON, ROBERT. 1973. The theory of rational option pricing. *Bell Journal of Economics and Management Science* 4 (Spring): 141–83.

NANDI, SAIKAT, 1998. How important is the correlation between returns and volatility in a stochastic volatility model? Empirical evidence from pricing and hedging in the S&P 500 index options market. *Journal of Banking and Finance* 22 (May): 589–610.

NANDI, SAIKAT, AND DANIEL F. WAGGONER. 2000. Issues in hedging option positions. Federal Reserve Bank of Atlanta *Economic Review* (First Quarter): 24–39.

RUBINSTEIN, MARK. 1994. Implied binomial trees. *Journal of Finance* 69 (July): 771–818.

Social Security in Latin America: Recent Reforms and Challenges

STEPHEN J. KAY AND BARBARA E. KRITZER

Kay is a senior economic analyst with the Latin America Research Group in the Atlanta Fed's research department. Kritzer is a senior research analyst with the Social Security Administration's Office of Policy.

The authors thank Tom Cunningham, Max Horlick, John Robertson, Paul Van de Water, and Martynas Ycas for helpful comments. The views expressed in this article are not necessarily those of the Federal Reserve Bank of Atlanta, the Federal Reserve System, or the Social Security Administration.

LATIN AMERICA HAS LED THE WORLD IN INTRODUCING INDIVIDUAL RETIREMENT SAVINGS ACCOUNTS INTENDED TO COMPLEMENT OR REPLACE DEFINED-BENEFIT STATE-SPONSORED PAY-AS-YOU-GO SOCIAL SECURITY PENSIONS. IN THE 1990S SEVERAL COUNTRIES IN THE REGION FOLLOWED CHILE'S LEAD IN SETTING UP INDIVIDUAL ACCOUNTS, AND SINCE THAT TIME COUNTRIES THROUGHOUT the world have looked to the region for lessons. This article describes the vast array of pension reforms that have taken place in Latin America since Chile's original 1981 reform and summarizes some of the fundamental policy challenges that remain. Policymakers seeking to learn from the Latin American reforms have no shortage of models from which to choose.

The first half of this article highlights the wide variety of policy choices that each country has made. Rather than presenting a comprehensive survey, this article describes some of the most noteworthy and unique features of each country's reform. While some countries have embraced defined-contribution individual accounts as a replacement for financially troubled state-run pension systems, other countries have adopted mixed systems or have made individual accounts optional and supplementary. The diversity of reforms in the region suggests that there is no single Latin American model but rather a range of pension systems which incorporate individual retirement savings accounts to varying degrees.

The article's second half describes some of the most serious policy challenges that policymakers have faced since the implementation of the new pension systems. Governments are seeking to improve the performance of the new systems by reducing administrative costs, limiting evasion, incorporating new categories of workers into the system, and improving competition in the pension fund industry. In this respect the region's diverse social security reforms are still works in progress.

The Range of Reforms in Latin America

The Chilean retirement pension scheme is used as a benchmark in this article and is described here in some detail, followed by brief descriptions of how other programs differ from the Chilean model.[1] While some countries have replaced their public social security system with a privatized one (Bolivia, El Salvador, and Mexico), others have added a new private tier and modified the public one (Argentina, Uruguay, Colombia, and Peru). Still others offer supplementary pensions (Costa Rica and Brazil). In describing the range of reforms in the region, countries are classified below according to whether or not workers are required to contribute to a private individual account (see Table 1). Selected reforms unique to each country are highlighted.

Mandatory Individual Accounts

In Chile, Bolivia, El Salvador, and Mexico, workers are required to contribute to individual retirement savings accounts, and the old public systems are either closed to new entrants or closed completely. Transition provisions may allow individuals who were in the labor force at the time of the reform to switch to the privatized system with some form of compensation.

Chile. In 1981 Chile was the first country to replace a pay-as-you-go system with mandatory individual retirement savings accounts. Prior to this reform, the Chilean pay-as-you-go system included inequitable benefits based on occupation and political clout, mismanaged programs, high rates of evasion, low coverage, and promises of higher benefits that could not be sustained. The pension reform was part

TABLE 1
Benefits under Privatized Systems

				Type of Retirement Available			
					Programmed		
	Retirement Age		Guaranteed		Withdrawals with	Programmed	Requirement for
Country	Men	Women	Minimum Pension	Annuity	Deferred Annuity	Withdrawals	Early Retirement
				Mandatory Individual Accounts			
Chile	65	60	Yes	Yes	Yes	Yes	Pension equals 50 percent of average wage over last ten years and 110 percent of minimum old-age pension.
Bolivia	65	65	Yes	Yes	No	No	Pension equals 70 percent of average of last five years' earnings.
El Salvador	60	55	Yes	Yes	Yes	Yes	Pension equals 70 percent of basic wage or 170 percent of minimum old-age pension.
				Mixed Systems			
Argentina	65	60	Yes	Yes	Yes	Yes	Pension equals 50 percent of average wage over last five years.
Uruguay	60	60	No	Yes	No	No	No early retirement.
Colombia	60[a]	55[b]	Yes	Yes	Yes	Yes	Pension equals 110 percent of minimum old-age pension.
Peru	65	65	No	Yes	Yes	Yes	Pension equals 50 percent of average salary in last ten years.
Mexico	65	60	Yes	Yes	Yes	Yes	Pension equals 30 percent of minimum old-age pension.

[a] Rising gradually to age 62 by 2014.
[b] Rising gradually to age 57 by 2014.

Source: SSA (1999)

of a major economic reform package that included privatization of state enterprises and government programs. Employees who moved from the old system to the privatized one received a government-mandated gross wage increase of 18 percent (about 11 percent net) to make up for the increase in the employee's contribution rate after the employer's contribution was eliminated (Ruiz-Tagle 1996, 3). Employees were also given a recognition bond representing the value of accrued rights under the old system that was indexed for inflation and funded by general revenues. To help fund the transition, Chile privatized state-owned industries and ran budget surpluses over a period of years (Diamond and Valdés-Prieto 1994, 280).

After the privatized scheme was introduced in 1981, the old system was closed to new entrants. All new workers were required to join the new system and those in the old system who were not within five years of retirement could opt to switch to the new system. The police and armed forces retained their separate programs.

Financing and Benefits. Under the new system, workers pay 10 percent of their monthly earnings into an individual retirement account run by a pension fund management company (*administradora de fondos de pensiones,* or AFP). The payment is mandatory for employees and voluntary for the self-employed. Workers pay a monthly administrative fee averaging 1.76 percent of salary, and an additional 0.64 percent of wages goes to survivors and disability insurance.[2] Most AFPs also charge a flat fee on contributions that averages about 600 pesos (about U.S.$1.15) (FIAP 2000b; SAFP 2000b). Workers may switch from one AFP to another twice per year without any transfer fee.

The retirement benefit is payable at age 65 (men) and 60 (women). Retirees choose among an annuity, programmed withdrawals (scheduled to guarantee income over the insured's expected lifespan), or a combination of the two options. The amount of the pension is based on the individual's contribution plus interest and less administrative fees. Pensions are protected against inflation because they are denominated in a monetary unit adjusted to reflect changes in the consumer price index—the *unidad de fomento*. Annuities are purchased from an insurance company for an additional administrative fee, which averages about 5.46 percent of the value of the annuity. Most AFPs also charge a monthly fee for programmed withdrawals that averages about 1.3 percent of the withdrawal; one AFP instead charges a flat monthly fee of 1,495 pesos (about U.S.$3) (SAFP 2000a, 2000b).

Early retirement is permitted if the pension equals at least 50 percent of the average wage over the last ten years and is at least 110 percent of the minimum old-age pension. A minimum pension is guaranteed to those who have made at least twenty years' worth of contributions and whose accumulated funds do not yield this minimum level (SSA 1999, 123).[3]

Table 2 compares key features of the financing of Chile's system with other systems in the region.

Administration. AFPs are private-sector companies strictly regulated as to allowable investments: as of August 2000, 37 percent of investments were in government bonds, 18 percent were in stocks, 33 percent were in the financial sector, and 12 percent were in international investments. Although the overall real rate of return on invested capital from July 1981 to April 2000 was about 11 percent, after subtracting the administrative fees mentioned above, the net rate of return was an average of about 7 percent (SAFP 2000c).

An AFP must maintain both a minimum and a maximum rate of return calculated to reflect the average performance of all AFPs. If an AFP exceeds the average by 2 percentage points or 50 percent (whichever is higher), it must place excess earnings in a rate of return fluctuation reserve fund.[4] An AFP must also keep 1 percent of the value of its pension fund as a separate reserve fund. Conversely, when returns are 2 percentage points below or 50 percent of the average (whichever is lower), an AFP must make up the difference from these reserve funds. If both of these funds become exhausted, the government makes up the difference, the pension fund management company is dissolved, and the individual accounts are transferred to another AFP.

Until recently, AFPs were permitted to manage only one fund. But since March 2000, AFPs have been required to offer a second fund that is completely

[1] For a more detailed description of the public and private retirement, disability, and survivors programs for all of these countries, see Social Security Administration (1999).

[2] Fees are assessed for each contribution made; thus, there are no fees for inactive accounts.

[3] The value of the minimum pension varies since it is adjusted when the consumer price index is at least 15 percent higher than in the previous year. It was 61 percent of minimum wage in 1982, 91 percent in 1987, and 74 percent in early 2000. Also, the minimum pension is about 7 percent higher for those over age 70 (CBO 1999; SAFP 2000a).

[4] Until 1999 the industry average was calculated yearly. Beginning in late 1999 the period was lengthened to three years but is being phased in; the period increases by one month every month for up to thirty-six months (SSB 1999, 131).

TABLE 2
Financing of Privatized Systems, December 1999

Country	Contribution (Percentage of Wages)		Average Survivors and Disability Insurance[a] (Percentage of Wages)	Average Administrative Fees[a] (Percentage of Wages)	Recognition Bonds
	Employee	Employer			
Mandatory Individual Accounts					
Chile	10	None	0.64	1.76[b]	Yes
Bolivia	10	None	2.0	0.50	Yes
El Salvador	3.25	6.75	1.13	2.05	Yes
Mexico	1.125	5.15	2.5[c]	1.79	No
Mixed Systems					
Argentina	11[d]	N/A[e]	1.01	2.40[b]	No
Uruguay	15[f]	N/A[e]	0.61	2.03	No
Colombia	3.375	10	1.86[c]	1.63[c]	Yes
Peru	10	None	1.38	2.36	Yes

[a] Employee pays as percent of earnings.
[b] Does not include flat fees.
[c] Employee and employer split fee.
[d] Administrative fee and survivors and disability insurance deducted from this amount.
[e] Employer's contribution goes to public system.
[f] For incomes above U.S.$800. Workers earning less than $800 contribute 7.5 percent of half of their earnings to an individual account and 7.5 percent of the other half to the public program.

Source: SSA (1999)

invested in fixed-rate instruments for workers within ten years of retirement (SSB 1999, 123).

The Superintendencia de Administradoras de Fondos de Pensiones (Superintendent of Pension Fund Management Companies, or SAFP), an autonomous state-financed government agency, regulates, supervises, and licenses AFPs. The superintendent, appointed by the president, also serves as a member of the risk-classifying commission that evaluates the risk of each type of allowable investment (Callund 1994, 409).

Bolivia. In addition to receiving retirement benefits by accumulating financial capital in personal accounts, all Bolivian resident citizens who were at least 21 years old on December 31, 1995, receive shares of 50 percent of the proceeds from the sale of six leading state enterprises. The original law called for a *bonosol*, an annual old-age bonus to be paid to all these citizens once they reach age 65. The assets are held in the *fondo de capitalizacion* (FCC) and invested and managed by AFPs. The *bonosol* also included funeral expenses (Von Gersdorff 1997, 10–12).

However, in 1998 the *bonosol* was suspended because the FCC was severely underfunded.[5] A new law replaced the *bonosol* with the *bolivida,* which is yet to be implemented. The *bolivida* is to be paid to Bolivians who were age 50 or older on December 31, 1995, when they reach age 65. It will be funded by 30 percent of the FCC. The remaining 70 percent of the FCC will finance an individual account (*cuenta de acciones populares*, or CAP) for Bolivians between the ages of 21 and 50 at the end of 1995. (The CAP is separate from the individual account financed by the employee's contribution.) Prior to retirement, an employee may use the CAP as collateral for a loan or to buy shares in the financial market; at retirement, he or she may convert the CAP to an annuity. The *bolivida* and CAP provisions will not be implemented until a new national identification system is in place because the old *bonosol* program had so many fraudulent claims.[6] The amount of the *bolivida* will be based on available funds (IMF 1998).

Currently, only two AFPs, both owned by a consortium of foreign firms, are permitted to operate. Enrollment is directed by the Bolivian government according to the enrollee's area of residence and date of birth. Since January 2000 enrollees have been permitted to switch AFPs if they have made twelve contributions, changed jobs, moved, or if fees or insurance premiums have increased. In 2005 other AFPs will be allowed to enter the market and enrollees will be permitted to switch from one AFP to another. Administrative fees are expected to increase at that time (SSB 1998, 88–105).

El Salvador. Eligibility for the privatized system is based on age. Men who were 55 or older and women who were 50 or older at the time of the 1998 reform were required to stay in the public ISSS (El Salvadorean Social Security Institute). Anyone under the age of 36 at the time of the reform was required to set up an individual account. Individuals between those ages could choose either system (SSA 1999, 117). Contribution rates are being phased in so that by the seventh year of operation (2005), employers will pay 8.75 percent of payroll, and employees will pay 4.25 percent of earnings plus 3 percent for survivors and disability insurance and administrative fees (SSA 1999, 117). The government will pay the minimum pension only if fiscal resources are available (Mesa-Lago 1997, 397).

As in Chile, AFPs must guarantee a rate of return that falls within a range of the average rate of return of all AFPs, but the government does not guarantee a minimum rate of return if an AFP goes bankrupt. In addition to investing in locally issued securities, AFPs invest in the state-run public housing fund for a period of ten years, beginning with 30 percent of total assets and gradually declining (Mesa-Lago 1997, 409; SSB 1998, 293).

Mexico. One unique feature of the reformed Mexican system is that workers may switch back from the private system prior to retirement. Instead of a recognition bond, retiring workers who have previously made contributions under the pay-as-you-go system may choose between a benefit under the old or new plan. If the individual chooses the pay-as-you-go benefit, the balance of his or her individual account is transferred to the government. This ability to choose is advantageous to the worker, who can receive the higher of the two benefits, but it poses a potential problem for the government, making it difficult to project and plan for the long-term costs of paying for the transition from the public to the privatized system. Having this option could cause workers to take higher risks with their individual accounts, increasing the financial burden if the resulting average account balances are low and the government must fund a large number of benefits under the old public system.[7]

A housing fund (INFONAVIT) sets up a separate interest-bearing housing account for each employee and is funded by a 5 percent payroll tax paid by the employer. Since the housing fund provides low-interest loans to employees for the purchase of a home, the returns from this account are lower than from the individual retirement account. Upon retirement, the balances of these two accounts are combined to provide the pension (Grandolini and Cerda 1998, 32–34; CBO 1999).

Each pension fund management company (*admisitrador de fondos de ahorro*, or AFORE) is limited to a 17 percent market share (to be increased to 20 percent after the system has been in existence for four years). Although each AFORE currently manages only one fund, in the future AFORES will be allowed to manage multiple funds with different types of investment and risk levels. AFOREs may charge fees for a variety of services. The fees are charges as a percentage of wages, as a percentage of assets under management (including inactive accounts), and as a percentage of real return (Grandolini and Cerda 1998, 19).

Unlike in many of the other Latin American countries where disability and survivors insurance is a separate private contract with an insurance company funded only by the employee, in Mexico the public Mexican Social Security Institute administers these programs. They are financed by an employer/employee/ government contribution (CBO 1999; SSB 1998, 202; 1999, 230; Queisser 1998, 92).

While some countries have embraced defined-contribution individual accounts as a replacement for financially troubled state-run pension systems, other countries have adopted mixed systems or have made individual accounts optional and supplementary.

Mixed Systems

While Chile, Bolivia, El Salvador, and Mexico developed systems that will eventually eliminate the state-run pay-as-you-go system and require all workers to contribute to private accounts, other countries kept their state-sponsored systems. The countries with mixed systems described below maintained their state-run programs and gave workers the option of contributing to private accounts. Some countries offer a first-tier state-provided benefit and the choice of a public or private benefit for the second tier; other countries allow switching from public schemes to private ones.

Argentina. The Argentine program has three tiers. The first two tiers are pay-as-you-go: a non-earnings-related universal flat-rate benefit based on

[5] Payments of the *bonosol* required the banks to borrow U.S.$50 million dollars at 11.75 percent interest per year. It became clear that in future years additional borrowing would be required (*La Razon* 2000).

[6] Recent newspaper accounts indicate that the program should begin during the first half of 2001.

[7] Although current regulations require most investments to be in government instruments, as the system matures, other types of investments will be permitted.

years of service and an earnings-related compensation benefit for service rendered before July 1994. The third tier offers a one-time choice between the public system and a private individual account (SSA 1999, 11). However, the Argentine pension system may soon undergo a major overhaul: in November 2000, President De la Rua announced a series of proposals that included eliminating both the universal flat-rate benefit and the state-run third-tier public benefit. At the time of publication, these proposed reforms had not been implemented.

A unique feature of the Argentine system is that a state-owned bank, the Banco de la Nación, is required to set up a pension fund management company that provides a guaranteed minimum rate of return equal to interest rates earned in savings accounts. The other pension fund administrators do not offer this kind of minimum guarantee; rather, they are expected to compete with the state-owned fund and provide returns that are equal or higher (Arenas de Mesa and Bertranou 1997, 334). Furthermore, the law states that no less than 20 percent of the investments of the Banco de la Nación must be invested in local economies; in practice, this requirement has called for investments in a combination of provincial, municipal, public sector enterprise, and autonomous public agency-issued bonds or Banco de la Nación–issued certificates of deposit.

The region's new pension systems continue to face a common set of policy dilemmas, including the need to improve finances, drive down expenses, reduce evasion, and expand coverage.

Uruguay. This program has a two-tier mixed system. The first tier covers all workers for the first U.S.$800 of monthly earnings (about 87 percent of the labor force earns under U.S.$800). The benefit is equal to a proportion of adjusted average monthly earnings. The second tier is a mandatory individual savings program for workers under the age of 40 with monthly earnings between approximately U.S.$800 and U.S.$2,400. (The program is voluntary for those who were age 40 when the program was set up and is voluntary for lower earners). At retirement, the insured must buy an annuity—indexed to average wages—from an insurance company (SSA 1999, 376).

Unlike most other Latin American countries where a new autonomous organization was created to oversee the private program, in Uruguay the central bank is responsible for the supervision and regulation of pension fund management companies. The social security bank supervises the public program and collects the contributions for both programs (Mesa-Lago 1997, 411; Mitchell 1996, 13).

Colombia. The Colombian system offers workers the choice between the public (Social Security Institute, or ISS) or private (AFP) retirement plans; workers are allowed to switch back and forth between the public and private plans every three years (SSA 1999, 82). Colombia's pension system is grappling with rapidly growing unfunded liabilities (Echeverry-Garzón and Navas-Ospina 1999, 93.) A government proposal in late 2000 to reduce these liabilities through a restructuring of the pension system does not appear to have sufficient political support.

By law, AFPs may offer more than one pension fund with different risk portfolios. Affiliates whose accounts would finance at least 50 percent of the minimum pension are permitted to invest the excess in other funds (Queisser 1997, 27).

Unlike in other countries where the supervisory institution is autonomous, in Colombia a government agency, the Superintendent of Banks, supervises both the public and private pension systems. It is effectively an umbrella regulatory body whose departments deal with banks, insurance companies, and pension funds. The superintendent's pension department regulates and supervises AFPs and other institutions and is not a separate and independent organization (Queisser 1998, 74; 1997, 26).

Peru. As an alternative to the existing pay-as-you-go system, Peru has introduced a private tier. Switching back and forth between the two tiers was permitted for the first two years of operation. Since then, once a worker has made a choice, no change is allowed. Peru is the only country that allows AFPs to charge an exit fee. In addition contributions to the AFPs are not tax-deductible and pensions are taxed (unlike in other countries); in effect, there is double taxation (Queisser 1997, 20). The Superintendent of Pension Fund Management Companies, an autono-mous agency that oversees the AFPs, is financed by 6.5 percent of gross earnings of all AFPs (AFP Horizonte 1997, 30).

Supplementary Accounts

While the countries described so far have initiated profound structural reforms that incorporate new systems of individual accounts, other countries have maintained their pay-as-you-go systems and provided optional supplementary accounts or private pensions. In Ecuador and the Dominican Republic, AFPs are operating even though proposals for reform of the pay-as-you-go system have been stalled in the legislature.

Costa Rica. Costa Rica's system differs from the countries described above in that the pay-as-you-go tier continues basically intact and voluntary individual accounts provide second-tier benefits. The pay-as-you-go benefit, financed by employee, employer, and government contributions, is equal to a proportion of adjusted average monthly earnings. The supplementary benefit is similar to the other programs described above.

A yet-to-be implemented new program reduces the employer's severance pay contribution by 3 percent. This 3 percent will go to a new funded labor account set up for each employee. Half of the account will fund the severance payment, and the other half will be sent to a supplementary pension fund chosen by the worker. Upon retirement individuals can choose either an annuity or programmed withdrawals (IBIS 1999, 39–40; 2000, 13). The voluntary account remains unchanged. The Superintendent of Pensions will oversee the new four-tiered social security system (IBIS 1999, 39–40).

Brazil. Brazil's private pensions are voluntary and serve as a supplement to the public system. Closed pension funds, the most common type (with about 92 percent of pension assets), are nonprofit and are set up by a company or group of companies, with membership limited to their employees. Both employees and employers contribute to these funds, and the funds are regulated by an agency within the Ministry of Social Security. Open pension funds are open to all workers and may be either nonprofit or for-profit organizations. The Superintendent of Private Insurance, a separate organization, oversees open pension funds.

In 1998 the Brazilian government introduced an individual programmed retirement fund (FAPI or *fundo de aposentadoria programada individual*) to supplement the public pension for workers who do not have the other private pension options. Workers choose an authorized financial institution or insurance company to manage their FAPI. Both employers and employees may contribute periodically (not less than once a year) for at least twelve months. After ten years, the employee may withdraw the funds in a lump sum or purchase some type of pension (SSB 1999, 106–14).

In some countries, where reform of the social security system is caught up in the legislative process, AFPs already operate despite the fact that no law has been passed. In Ecuador, six AFPs with a total of about 200,000 affiliates have about U.S.$35 million in assets. The Dominican Republic has four AFPs with about 15,000 affiliates and about U.S.$75 million in funds (FIAP 2000a, 36–38; SSB 1999, 290).

Policy Challenges

Throughout the region, the increasing prevalence of defined-contribution individual retirement savings programs is reflected in the increase in pension fund investments as a percentage of GDP (see Table 3). A new set of policy challenges has accompanied this broad set of reforms, including high commission costs, limited competition within the pension fund industry, questions over investment rules, high evasion rates, greater differentiation in pensions based on gender, and political obstacles to incorporating occupational groups not currently participating in the new system. These policy issues, and the extent to which countries in the region have attempted to resolve them, are summarized below. The discussion concentrates on Chile's system of individual accounts, established in 1981, because its relative maturity

TABLE 3
Pension Fund Management Firm Assets as a Percentage of GDP, 1999

Country	Year Program Began	Company Assets (in Thousands of U.S.$)	Fund as Percentage of GDP
Argentina	1994	16,787,099	5
Bolivia	1997	534,803	6
Chile	1981	34,501,000	47
Colombia	1994	2,887,108	5
El Salvador	1998	212,591	2
Mexico	1997	11,508,822	2
Peru	1995	2,406,034	4
Uruguay	1996	591,161	2

Source: FIAP (2000a; 2001)

has led to challenges that the other, newer systems, may not yet have encountered. (The second-oldest system began twelve years later in Colombia).

Administrative Fees. Under the new systems of individual accounts, workers contribute a percentage of their wages to their individual accounts, with additional deductions going toward an administrative fee and insurance premium. Commission fees vary throughout the region. In Colombia, administrative fees paid to pension fund administrators represent 14 percent of total contributions (not including insurance); the figures in Uruguay, Peru, and El Salvador are 21.2 percent, 22.8 percent, and 31.3 percent, respectively (calculations based on FIAP 2000b). Some argue that high administrative costs may be a necessary feature of a "retail competition" model in which pension fund administrators compete directly for worker contributions (Thompson 1999, 9).

In Argentina an average of 24 percent of a worker's total contributions goes toward an administrative fee, and in Chile the figure is 15 percent (FIAP 2000b). These two countries also allow firms to charge an additional flat-rate commission fee. In Argentina fees range from $1.90 to $9.00 for the eight firms out of twelve that charge a flat-rate fee (SAFJP 2000) whereas in Chile the charges range from 73 cents to $1.89 (only one of the eight pension firms in Chile does not charge such a fee) (SAFP 2000b). The net result is that lower-income workers pay a higher percentage of their salaries in charges than do higher income workers. For example, an Argentine worker earning $240 a month would pay an average of 3.99 percent of salary in total charges (commission fees plus disability insurance) while a worker earning $2,400 a month would pay 3.19 percent (SAFJP 2000).

Commission fees have a significant impact on returns, especially in the early years of a new system, when workers are beginning to accumulate capital in their accounts. In Chile the return on capital between July 1981 and April 2000 was 11.1 percent, but once commissions are factored in, lower-income earners received a 7.34 percent return, and higher-income earners received a 7.69 percent real average return (SAFP 2000c).[8] When workers retire, they may purchase annuities or elect to withdraw their money gradually in a programmed withdrawal. In Chile, annuities are purchased from an insurance company, and in April 2000 average fees were 5.46 percent of the value of the annuity (SAFP 2000a).

Policymakers have made lowering administrative fees a top priority. Pension fund administrators in Argentina have considered dropping fixed commissions altogether (Ojeda 2000). In Chile the government has sought to promote greater consumer awareness by requiring pension funds to publish data on expenses and fees. The government also made the process of transferring from one pension fund to another slightly more cumbersome (instead of simply signing a form, workers were required to present identification and a recent account statement). As a net result, transfers dropped from 26 percent of the labor force in 1997 to 3.5 percent in June 1999, enabling pension funds to reduce sales and marketing expenses; the pension fund sales force dropped from 22,643 employees in November 1997 to 4,026 in September 1999 (SSB 1999, 130; *Santiago Times* 1999). During that time the variable commission fee fell: between November 1997 and July 2000 average administrative fees dropped from around 19 percent to 15 percent of total contributions. However, the average flat fee rose 37 percent during the same period (SAFP 1997–2000). Several other reforms, such as loyalty discounts for workers who remain with a fund for a certain length of time, group discounts, and commission fees that vary according to services provided, are also under discussion in Chile.

Limited Competition. Numerous observers have attributed the high expenses of private pension funds to limited competition and industry concentration within the pension fund industry. The combination of a small market and the economies of scale inherent to pension fund management may create a tendency toward oligopoly, and concentration within the industry may limit the extent to which market competition can drive down costs (Thompson 1999, 27).

Diamond and Valdes-Prieto (1994, 288) have noted that the Chilean AFP market resembles a monopolistic competitive market rather than a competitive market, preventing lower costs and returns on the risk-return frontier. At its peak, there were twenty-two pension fund companies in Chile, a number which now stands at eight due to consolidation. The largest three funds control over two-thirds of the market, and, in an effort to stimulate competition and lower costs, the Chilean government now allows pension funds to subcontract investment services with other financial services firms.

In other parts of the region, Argentina began with twenty-six pension funds and now has thirteen. Mexico has gone from seventeen to thirteen funds. Two of Colombia's eight firms control almost 50 percent of the market. One of Uruguay's six firms controls 55 percent (FIAP 2000c). See Table 4 for a further comparison of the number and characteristics of pension fund management firms.

There has been much debate over whether or not banks, mutual funds, and insurance companies

TABLE 4
Characteristics of Pension Fund Management Firms

Country	Name	Number of Companies	Allowable Funds per Company	Allowable Transfers per Year	Minimum Rate of Return
		Mandatory Individual Accounts			
Chile	AFP	8	Two	Twice	Yes
Bolivia	AFP	2	One	Once	Yes
El Salvador	AFP	5	One	Every 1.5 years	Yes
Mexico	AFORE	13	One[a]	Once	No
		Mixed Systems			
Argentina	AFJP	13	One	Twice	Yes
Uruguay	AFAP	6	One	Twice	Yes
Colombia	AFP	8	Multiple	Twice	Yes
Peru	AFP	5	One	Twice	Yes

[a] The law allows for multiple funds; however, there are no regulations at this time.

should be allowed to enter the pension fund market to compete directly with AFPs. These firms argue that greater competition would spur competition and lower costs. Pension fund companies caution that allowing these firms entry could lead to conflicts of interest. At this point the subject is still being debated in Chile, but the government has signaled that it is committed to stimulating greater competition within the pension fund industry.

Investment Rules. Pension fund investments are strictly regulated, and pension funds generally face limits regarding what percentage of their portfolios can be invested in a given type of security. Local capital markets suffer from a lack of diversification of investment-grade instruments and are dominated by the issuance of government paper; therefore, investment is highly concentrated in state-issued bonds and short-term instruments. Uthoff (1997) argues that these factors tend to result in less-than-efficient allocation of investment and long-term capital formation.

Because pension funds receive sanctions for deviating from the average return, there is a herd effect as firms have little incentive to take risks that would lead to deviation from the mean. This lack of incentive effectively rules out longer-term investment strategies. Beginning in October 1999, Chile began lengthening the period of time for calculating minimum profitability from one to three years. Peru relaxed its minimal profitability requirements as well and began calculating minimum returns over a five-year period. These initiatives may encourage longer-term investment strategies and reduce the herd effect.

Until recently each Chilean pension fund company offered only one portfolio for all workers despite the fact that workers have a range of preferences for risk taking that can vary according to age and proximity to retirement. Following a severe financial market downturn in Chile in 1998, many workers decided to delay retirement. The government later approved the creation of a second type of investment portfolio to be invested only in fixed-income instruments and available to men aged 55 or older and women aged 50 or older. The pension fund superintendency is also studying the possibility of adding a third pension fund geared toward younger workers at the opposite end of the risk/return spectrum, which would be invested largely in stocks.

As the pioneer of private pension funds in the region, Chile has gradually liberalized investment rules as the system has matured and become more established. For example, no foreign investment was permitted in the early years of the system, but currently up to 20 percent of pension fund investments may go overseas. Recently introduced legislation would raise that limit to 35 percent, reducing risk through greater diversification.

Mexico's AFOREs currently manage only one pension fund; however, in the future the law will permit each AFORE to offer funds with different types of investments and risk levels. Each AFORE will be required to offer one fund that has at least 51 percent of its investments in inflation-indexed securities, one fund with mainly fixed-income investments, and another fund with investments primarily in equities (Queisser 1998, 92). Colombia permits pension funds to operate more than one plan, and affiliates whose

[8] A study by CB Capitales (1999) concluded that once commission fees are accounted for, the real average return from 1981 through 1998 for a worker earning the average wage was 5.1 percent.

funds would finance at least 50 percent of the minimum pension are permitted to invest the excess in other plans.

Evasion. As was the case with the public systems, evasion has remained a consistent problem since the inception of new private pension systems. Because the informal sector in Latin America comprises about 57 percent of the labor force (Lora and Olivera 1998, 7–8), by definition less than half the workforce is in a position to make social security contributions. The growth in the informal sector shows no sign of abating. For example, approximately 72 percent of new employment in Argentina in the 1990s was in the informal sector (Ojeda 2000). Only about half of those workers with a private pension plan make regular contributions. Therefore, only about a quarter of the total workforce is making regular contributions and is on track to receive full retirement benefits. Workers who do not make regular contributions will, of course, accumulate less capital in their individual accounts and will receive lower retirement pensions. (At the end of 1999, 61 percent of workers in Mexico with individual pension accounts made regular contributions, 56 percent in Uruguay, 44 percent in Peru, Argentina, and Chile, and 40 percent in Colombia [FIAP 2000d]).

Pension reforms are continually subject to revision, and reform itself can be an incremental process. Latin America's social security systems are likely to continue to attract international attention as they confront the ongoing challenges of reform.

In some cases, there are incentives to evade. For example, in Chile workers who contribute for twenty years receive a government subsidy that will top up their benefit and assure them a minimum pension. Workers therefore have little incentive to contribute over and above the twenty-year requirement. In Argentina workers must contribute for thirty years to be eligible for a full benefit—a daunting obstacle in a country with high unemployment and a large informal sector. If governments do not improve compliance, significant percentages of the workforce will likely require government subsidies (if available) in order to receive adequate retirement benefits. The issue of evasion is closely tied to the overall structure of regional labor markets and is unlikely to improve until the size of the informal sector begins to shrink.

Gender. The new systems of individual accounts in Latin America strictly link benefits with earnings and place men and women in separate actuarial categories. Under the old systems, women and men were placed in the same actuarial category, and differences in pension levels were less pronounced since benefits did not depend directly upon total contributions and investment results. Because women tend to earn less than men and spend more years of their lives outside the paid labor force, women will also tend to accumulate less capital than men. With men and women now placed in separate actuarial categories, a woman purchasing an annuity and having the same amount of money as a man will receive lower benefits than the man because of her greater expected longevity.

According to a 1995 Chilean study, a woman whose salary is 75 percent of a given man's salary would receive a pension that is between 35 percent and 45 percent of his pension. If a woman and a man have the same salary and have contributed for the same number of years, the woman's pension would be between 52 percent and 76 percent of the man's pension (though she could expect to collect it longer). Consequently, in order for a woman to receive the same pension as a man with the same salary, she must retire later than the man does (Arenas de Mesa and Montecinos 1999). Thus far about half of those who have retired in Chile under the new system have accumulated sufficient funds to retire early. Of that total, 86 percent have been men and only 14 percent have been women (SAFP 1997–2000).

Incorporating New Sets of Workers. Pension reforms in the region often exclude workers benefiting from special, privileged pensions. The military and police, for example, have generally not been included in any of the new individual accounts systems in the region (Bolivia is one exception). In Colombia the new private system does not cover the military, national police, teachers, or employees of the state-owned oil company. In Argentina and Uruguay reforms were intended to eventually include the military and police, but despite some discussion little progress has been made. The 1995 Uruguayan reform legislation stated that programs would be introduced by the end of 1996 to incorporate workers enrolled in the relatively generous quasi-governmental pension plans into the new system of individual accounts (such systems include pension programs for bank employees, notaries, and professionals). However, these occupational groups have continued to object to being incorporated into the new system, largely because their current pension plans grant them a defined-benefit pension that is likely to be higher than what they would receive in a defined-contribution individual account plan.

For this reason, governments seeking to incorporate additional occupational groups will continue to face political opposition.

Argentina and Uruguay are the only countries where affiliation is mandatory for the self-employed. In other countries affiliation is optional. In 1997 only 11 percent of Chile's self-employed were affiliated with a pension fund, and only 4 percent made regular contributions (SAFP 1998, 196). It is clear that providing adequate pension coverage for the self-employed will be an ongoing challenge in the region.

Conclusion

Latin American countries have become the world's laboratory for pension systems based upon individual retirement savings accounts. This article illustrates the broad range of pension reforms by highlighting features specific to each country. Chile, Mexico, and El Salvador have gone the farthest in converting to individual accounts; Argentina, Uruguay, Colombia, and Peru have allowed parallel pay-as-you-go systems to continue. Under Bolivia's reform, retirement accounts were capitalized with revenue from the sale of state-owned assets. Meanwhile, several countries, including Brazil and Costa Rica, have embraced voluntary, supplemental pension plans but remain committed to the public pay-as-you-go system. Clearly there is no single Latin America model, and countries that seek to learn from the Latin American reforms can look to a wide range of systems.

The process of pension reform in Latin America remains a work-in-progress as initial reforms have been revised to reflect new policy needs. The region's new systems continue to face a common set of policy dilemmas, including the need to improve finances, drive down expenses, reduce evasion, and expand coverage. In Chile, for example, regulatory restrictions continue to evolve after twenty years. Argentina's president recently announced that he favors ending the current state-sponsored universal pay-as-you-go benefit and replacing it with a residual, means-tested benefit. Colombia's government also intends to revamp its troubled pension system in order to cope with unsustainable benefit obligations. These developments serve as a reminder that pension reforms are continually subject to revision, and that reform itself can be an incremental process. Latin America's social security systems are likely to continue to attract international attention as they confront the ongoing challenges of reform.

REFERENCES

AFP Horizonte. 1997. *Tres años del sistema privado de pensiones*, 1993–1996. Peru: AFP Horizonte.

Arenas de Mesa, Alberto, and Fabio Bertranou. 1997. Learning from social security reforms: Two different cases, Chile and Argentina. *World Development* 25 (March): 329–42.

Arenas de Mesa, Alberto, and Veronica Montecinos. 1999. The privatization of social security and women's welfare: Gender effects of the Chilean reform. *Latin American Research Review* 34 (3): 7–37.

Callund, Jonathan D.H. 1994. Description of the social security system in Chile. *Assurances* 3 (October): 391–415.

CB Capitales. 1999. *Comentario macroeconómico, primera quincena de abril de 1999*. Chile: CB Capitales.

Congressional Budget Office (CBO). 1999. *Social security privatization: Experiences abroad.* January.

Diamond, Peter, and Salvador Valdés-Prieto. 1994. Social security reforms. In *The Chilean economy: Policy lessons and challenges*, edited by Barry P. Bosworth, Rudiger Dornbusch, and Raúl Labán. Washington, D.C.: Brookings Institute.

Echeverry-Garzón, Juan Carlos, and Verónica Navas-Ospina. 1999. Confronting fiscal imbalances via intertemporal economics, politics, and justice: The case of Colombia. In *Sustainable public sector finance in Latin America: A conference presented by the Latin America Research Group.* Atlanta: Federal Reserve Bank of Atlanta.

Federación Internacional de Administradoras de Fondos de Pensiones (FIAP). 2000a. *Bulletin Number 7*. March.

———. 2000b. *Estructura de Comisiones*. <http://www.fiap.cl> (September 20, 2000).

———. 2000c. *Fondos Administrativos*. <http://www.fiap.cl> (September 15, 2000).

———. 2000d. *Anexo I: Número de Afiliados, Fondos Administrados, Composición de la Cartera por país al 31.12.99*. <http://www.fiap.cl> (November 17, 2000).

———. 2001. *Informacíon Económica Relevante Por País.* <http://www.fiap.cl> (February 22, 2001).

Grandolini, Gloria, and Luis Cerda. 1998. The 1997 pension reform in Mexico. World Bank Policy Research Working Paper No. 1933, June.

International Benefits Information Service (IBIS). 1999. IBIS Briefing Service, November.

———. 2000. IBIS Briefing Service, January.

International Monetary Fund (IMF). 1998. *Bolivia: Memorandum of economic policies.* Washington, D.C.: International Monetary Fund. <http://www.imf.org/external/np/loi/081498.htm> (October 31, 2000).

La Razon (La Paz, Bolivia). 2000. Las AFP tienen $U.S.13 millones para el Bolivida. August 2.

Lora, Eduardo, and Mauricio Olivera. 1998. Macro policy and employment problems in Latin America. Inter-American Development Bank Working Paper 372, March.

Mesa-Lago, Carmelo. 1997. Comparative analysis of structural pension reform in eight Latin American countries: Description, evaluation, and lessons. In *Capitalization: The Bolivian model of social and economic reform,* edited by Margaret Pearce. Washington, D.C.: Woodrow Wilson Center, Current Studies on Latin America.

Mitchell, Olivia. 1996. Social security reform in Uruguay: An economic assessment. Pension Research Council Working Paper No. 96-20.

Ojeda, Laura Luz. 2000. Estudian quitar los cargos fijos para bajar los costos de las AFJP. *La Nacíon* (Buenos Aires, Argentina), June 29.

Queisser, Monika. 1997. Pension reform and private pension funds in Peru and Colombia. The World Bank Financial Sector Development Department Policy Research Working Paper No. 1853, November.

———. 1998. The second-generation pension reforms in Latin America. Organisation for Economic Co-operation and Development Maintaining Prosperity in an Aging Society Working Paper AWP5.4.

Ruiz-Tagle, P. Jaime. 1996. El nuevo sistema de pensiones en Chile: Una evaluación provisoria (1981–1995). Programa de economía del trabajo materia de discusión no. 13. Santiago, Chile (January).

Salomon Smith Barney (SSB). 1998. Private pension funds in Latin America—1998 update. Latin America Equity Research Industry Report, December.

———. 1999. Private pension funds in Latin America—1999 update. Latin America Equity Research Industry Report, December.

Santiago Times (Santiago, Chile). 1999. Transfers between pension funds drop 75 percent. July 28.

Social Security Administration Office of Policy (SSA). 1999. *Social security programs throughout the world—1999.*

Superintendencia de Administradoras de Fondos de Jubilaciones y Pensiones (SAFJP). 2000. *Memoria Trimestral*. Various issues.

Superintendencia de Administradoras de Fondos de Pensiones (SAFP). 1997–2000. *Boletín Estadístico.* Various issues.

———. 1998. *El Sistema Chileno de Pensiones, Cuarta Edición*, April.

———. 2000a. *Boletín Estadístico* 155 (April).

———. 2000b. *Estructura de Comisiones del Fondo Tipo 1.* <http://www.safp.cl> (October 20, 2000).

———. 2000c. *Rentabilidad Real Annual de la Cuenta de Capitalizacíon Individual.* <http://www.safp.cl> (October 5, 2000).

Thompson, Lawrence H. 1999. Administering individual accounts in social security: The role of values and objectives in shaping options. The Urban Institute Retirement Project Occasional Paper No. 1.

Uthoff, Andras. 1997. Reforma de los sistemas de pensiones y ahorro. In *Ahorro nacional: La clave para un desarrollo sostenible*. Madrid: IRELA.

Von Gersdorff, Hermann. 1997. The Bolivian pension reform: Innovative solutions to common problems. The World Bank Financial Sector Development Department Working Paper No. 1832. <http://econ.worldbank.org/docs/755.pdf> (February 2, 2002).