WILLIAM F. SHARPE*

Capital Asset Prices with and without Negative Holdings

Introduction

Following tradition, I deal here with the Capital Asset Pricing Model, a subject with which I have been associated for over 25 years, and which the Royal Swedish Academy of Sciences has cited in honoring me with the award of the Prize in Economic Sciences in Memory of Alfred Nobel.

I first present the Capital Asset Pricing Model (hence, CAPM), incorporating not only my own contributions [1] but also the outstanding work of LINTNER (1965, 1969) and the contributions of MOSSIN (1966) and others. My goal is to do so succinctly yet in a manner designed to emphasize the economic content of the theory.

Following this, I modify the model to reflect an extreme case of an institutional arrangement that can preclude investors from choosing fully optimal portfolios. In particular, I assume that investors are unable to take negative positions in assets. For this

version of the model I draw heavily from papers by GLENN (1976), LEVY (1978), MERTON (1987) and MARKOWITZ (1987, 1990).

Finally, I discuss the stock index futures contract - a major financial innovation of worldwide importance that postdates the development of the CAPM. Such contracts can increase the efficiency of capital markets in many ways. In particular, they can bring actual markets closer to the idealized world assumed by the Capital Asset Pricing Model.

The Capital Asset Pricing Model

The initial version of the CAPM, developed over 25 years ago, was extremely parsimonious. It dealt with the central aspects of equilibrium in capital markets and assumed away many important aspects of such markets as they existed at the time. In the last 25 years, theorists have extended arid adapted the approach to incorporate some of these real-world phenomena. Important examples are LINTNER's (1969) version, which focuses on returns in real terms; BRENNAN's (1970) version, which deals with the effects of taxation; BLACK's (1972) version in which there is no riskless asset; MERTON's (1973) version, which deals with a more general class of utility functions; KRAUS and LITZENBERGER's (1976) version, which takes into account the third moment of the return distribution; LEVY's (1978) version, which incorporates in-

vestors' concern with future investment opportunities; RUBINSTEIN's (1974) version, which incorporates transactions costs; BREEDEN's (1979) version, which deals with market segmentation; and MARKOWITZ' (1990) version, which considers restrictions on short sales.

Throughout, empiricists have subjected variations of the model to tests of increasing power. Moreover, alternative models have been proposed, most notably ROSS' (1976) Arbitrage Pricing Theory.

While theorists have been busy adapting the CAPM to incorporate real-world impediments to efficiency, practitioners have been equally busy reducing some of those impediments. Many of the financial instruments and institutions developed in the last decade serve to better "complete the markets" - in particular, to allow a more efficient distribution of risk among investors.

I will not attempt a general treatment of the institutional costs and constraints that can affect the efficiency with which risk is allocated, nor of the many recent financial innovations that serve to reduce such costs and constraints. Instead, I will focus on one prototypical example - restrictions on negative positions in securities, and one such innovation - stock index futures.

To facilitate the exposition, many formulas and proofs will be relegated to footnotes. In this I follow a long personal tradition, since a key proof in SHARPE (1964) - my first published paper on Capital Asset Pricing Model - was contained in a footnote. [2]

The CAPM and Financial Economic Theory

A common taxonomy of work in financial economics differentiates between *normative* (prescriptive) and *positive* (descriptive) theories. MARKOWITZ' (1952) path-breaking mean-variance portfolio theory falls cleanly into the former category, dealing as it does with rules for optimal portfolio choice by an individual. The CAPM can be neatly classified as belonging to the latter, since it is concerned with the determination of the prices of capital assets in a competitive market. But two such categories are not sufficient. Much of the work in the field can best be described as approaching *normative issues in a positive context*. The seminal MODIGLIANI/MILLER (1958) model is of this genre, since it prescribes optimal behavior of a corporation faced with a capital market in which security prices are determined by the actions of individual investors cognizant of opportunities for substitution.

Even this three-way taxonomy fails to capture the interrelationships among the alternative approaches. Most positive models in financial economics, like those in the broader field of economics, are built on normative foundations. Individuals engaging in maximizing behavior are assumed to interact with one another until an equilibrium condition is reached. This is clearly the case with the CAPM, which explicitly assumes that investors follow the prescriptions of Markowitz' portfolio theory. Moreover, as in traditional microeconomic theory, financial economic theories of equilibrium relationships are taken as prescriptions for decision in markets that may not strictly conform to the conditions of the theory.

The domain of positive financial economics theory is sometimes divided into a set of models that may be termed *utility-based* and a complementary set that may be termed *arbitrage-based*. Models that fall in the latter category derive implications from the assumption that capital asset prices will adjust until it is impossible to find a strategy that requires no initial investment, provides a positive cash flow in at least one state of nature, and requires no negative cash flow in any future state of nature. [3] Models that fall in the former category typically conform to the conditions required for the latter, but derive stronger implications due to added assumptions about the *utility functions* that investors are assumed to maximize. [4]

Much of the early work in financial economics dealt with markets in which the interaction of a large number of individuals, each equally informed, determined prices. In this sense the work followed the tradition of competitive equilibrium theory in economics. More recently, attention has focused on markets in which there are few participants and/or in

which different individuals have different sets of information.

The CAPM is, of course, a theory in the earlier tradition of the field. It is a positive theory, incorporates assumptions about investors' utility functions, and assumes a market with a large number of participants, each of whom has access to the same set of information.

Capital Asset Prices with Negative Holdings

Assumptions

Assume that the economy consists of K investors. Investor k's invested wealth, expressed as a proportion of the total wealth invested by all investors is W_k. He or she wishes to maximize:

$$U_k = E_k - \frac{V_k}{\tau_k} \tag{1}$$

For expository convenience, I will term U_k investor k's *utility*. It may be regarded as a primitive utility function. Alternatively, it may be considered an approximation to the investor's *expected utility* in the sense of VON NEUMANN and MORGENSTERN (1944). If the investor is assumed to have a negative exponential utility function over wealth and returns are jointly normally distributed, the approximation will be exact. Even if the investor has some other utility function and/or returns are not jointly normally distributed, U_k may provide an excellent approximation, as shown by LEVY and MARKOWITZ (1979).

In equation (1), E_k is the expected return on investors k's portfolio, V_k is the variance of the portfolio, and τ_k is his or her risk tolerance. Investors differ in risk tolerance.

This relationship can be interpreted in a number of useful ways. Clearly, risk tolerance measures an investor's *marginal rate of substitution of variance for expected return*. For convenience we assume that each investor's risk tolerance is constant over the feasible range of expected return and variance.[5] The value obtained by dividing portfolio variance by an investor's risk tolerance can be characterized

as a *risk penalty*, leading to the interpretation of U_k as a *risk-adjusted expected return*. Alternatively, U_k may be considered a *certainty-equivalent* return, since a portfolio with a return of U_k and zero risk would have the same utility for the investor as the portfolio in question.

This is the objective function of MARKOWITZ' (1952) "mean-variance" approach to portfolio selection. It is a highly parsimonious characterization of investors' goals, employing a myopic view (i.e. "one period at a time") and focusing on only two aspects of the probability distribution of possible returns over that period. [6] Moreover, it assumes that the investor [7] can assess at least the first two moments of the probability distributions associated with alternative investment portfolios. The genius of the approach is its ability to capture much of what matters to investors. Moreover, it serves well as a base for extensions and adaptations designed to accommodate additional aspects of investor's preferences.

As in Markowitz' work, the expected return on a portfolio depends on the expected returns on its component securities. [8] A portfolio's risk depends on both the risks of the component securities and on their correlations with one another. More succinctly: portfolio risk depends on the covariances among securities. [9]

All investors are in agreement concerning expected returns and covariances. Of course, these moments of the joint distribution of security returns will be a function of security prices. When equilibrium prices are attained, however, each investor will choose an optimal portfolio, given the current values of expected returns and covariances; moreover the resultant portfolio choices will cause the markets to clear. Investors are allowed to take negative positions in one or more assets. Thus holdings may be positive, zero or negative. There are no transactions costs or other constraints and asset positions are fully divisible.

Portfolio optimality

Investor k seeks to maximize U_k subject to a *full investment constraint* of the form:

$$\sum_i X_{ik} = 1$$

where X_{ik} represents the proportion of investor k's portfolio invested in asset i.

To do so, he or she must select a portfolio in which the *marginal utility* of every security is the same. If this were not the case, it would be possible to reallocate funds from a security with a lower marginal utility to one with a higher one, thereby increasing utility without violating the full investment constraint.

This *first order condition* for portfolio optimality can be represented as follows:

$$E_i - \frac{2}{\tau_k} C_{ik} = \lambda_{fk} \qquad \text{for all } i. \tag{2}$$

Here C_{ik} is the covariance of security i with investor k's optimal portfolio and λ_{fk} is investor k's *marginal utility of wealth.* [10]

Aggregation

Assume that markets have cleared, so that all securities are held by the K investors in the economy. Relationships among key variables can be examined by aggregating the conditions that must hold when each investor obtains an optimal solution, taking into account the relative amounts of wealth that each has invested. To do so requires only a few straightforward operations. In effect, a wealth-weighted average is taken of the first-order conditions for each security. Not surprisingly, the result is similar to that obtained earlier.

$$E_i - \frac{2}{\tau_m} C_{im} = \lambda_{fm} \qquad \text{for all } i. \tag{3}$$

Here, τ_m is the wealth-weighted risk tolerance of the investors in the market, or the *societal risk tolerance.* The value of C_{im} represents the covariance of security i with the *market portfolio*, which includes all securities in the market, with each represented in proportion to its outstanding value. The last term is a weighted average of the values of λ_{fk} for the K investors, with the weights depending on the investors' influences in the marketplace, where influence depends on both invested wealth and risk tol-

erance. It can be interpreted as the *societal marginal utility of wealth.* [11]

Expected returns

One of the key implications of the CAPM concerns the relationship among the expected returns of capital assets. It can be obtained by a minor rearrangement of the previous equation to give:

$$E_i = \lambda_{fm} + \frac{2}{\tau_m} C_{im} \qquad \text{for all } i. \tag{4}$$

This shows that in equilibrium there is a linear relationship between the expected returns on securities and their covariances with the market portfolio. Usually the relationship is expressed in terms of a security's *beta*, a scaled measure obtained by dividing a security's covariance with the market portfolio by the variance of market portfolio (V_m). Substituting this measure gives:

$$E_i = \lambda_{fm} + \frac{2V_m}{\tau_m} \beta_{im} \qquad \text{for all } i, \tag{5}$$

where:

$$\beta_{im} \equiv \frac{C_{im}}{V_m}$$

Of course, equation (5) is also linear. Moreover, since portfolio expected returns and covariances with the market portfolio are simply value-weighted averages of the corresponding measures for the component securities, it follows that this relationship holds for all portfolios as well as for all securities.

The risk premium

Since the previous equation holds for any portfolio, it will hold for the market portfolio itself. Moreover, the beta value of the market portfolio with itself must equal 1. Letting E_m represent the expected return on the market portfolio, these relationship imply that:

$$\frac{E_m - \lambda_{fm}}{V_m} = \frac{2}{\tau_m} \tag{6}$$

The term on the left-hand side is the *risk premium per unit of variance*. As the equation shows, it is inversely related to societal risk tolerance.

The Security Market Line

Substituting the risk premium per unit of variance for $2/\tau_m$ in the equation for security expected returns gives a more traditional form of the relationship:

$$E_i = \lambda_{fm} + (E_m - \lambda_{fm})\beta_{im} \qquad \text{for all } i. \qquad (7)$$

A graphical portrayal is termed the *Security Market Line*. The CAPM implies that all securities and portfolios will plot along such a line. Many would argue that this relationship is the most important single conclusion derived from the CAPM. It shows that expected returns will be linearly related to *market risk*, but no, as often believed, to *total risk*.

Riskless borrowing and lending

As can be seen from equation (7), λ_{fm} may be interpreted as the expected return on any "zero-beta" portfolio, including the zero-beta portfolio with minimum variance, as suggested by BLACK (1972). If a riskless asset is available, λ_{fm} will equal R_f, the riskless rate of return. Under these conditions, the Security Market Line relationship can be written:

$$E_i = R_f + (E_m - R_f)\beta_{im} \qquad \text{for all } i. \qquad (8)$$

Henceforth, we assume that a riskless asset does exist and can be held in positive or negative amounts - i.e., that investors may either lend or borrow at the riskless rate R_f.

The characteristic line

The value of β_{im} may be given an interpretation similar to that found in regression analysis utilizing historic data, although in the context of the CAPM it is to be interpreted strictly as an ex ante value based on probabilistic beliefs about future outcomes. The relationship between $\tilde{R}_i$ and $\tilde{R}_m$, the stochastic returns on security i and the market portfolio, respectively, can be written as:

$$\tilde{R}_i = a_i + \beta_{im}\tilde{R}_m + \tilde{\varepsilon}_i \qquad (9)$$

Given the manner in which β_{im} is defined, it must be the case that $\tilde{\varepsilon}_m$ is uncorrelated with [12] $\tilde{R}_m$. Moreover, a_i can be defined so that the expected value of $\tilde{\varepsilon}_i$ is zero. However, there is no reason to expect that $\tilde{\varepsilon}_i$ - the residual return or non-market component of return for security i - will be uncorrelated with the comparable component for security j. [13] While the CAPM places no restrictions on the correlation of the residual terms, it does restrict the values of the intercept (a_i) terms. Since the expected value of ε_i is zero, the Security Market Line relationship requires that each intercept is related directly to the security's beta value. [14] The CAPM thus implies that:

$$\tilde{R}_i = (1 - \beta_{im})R_f + \beta_{im}\tilde{R}_m + \tilde{\varepsilon}_i \qquad (10)$$

A graphical portrayal of this relationship is termed a security or portfolio's *characteristic line*.

Factor models of security returns

Much confusion has arisen concerning the relationship between the equilibrium results of the CAPM and the underlying relationships among security returns. As can be seen, the CAPM makes *no* assumptions about the "return generating process". Hence, its results are completely consistent with *any* such process.

Early approaches to portfolio selection [15], assumed that returns were generated by a model similar (but not identical) to that of equation (9), with the further condition that the residual values were uncorrelated across securities. [16] My initial approach to capital asset pricing in SHARPE (1961) made a similar assumption. Such a "single index" or "single factor" model represents a special case of a *factor model of security returns*. Multi-factor models have been explored by a number of researchers and currently enjoy widespread use in financial practice.

A factor model of security returns identifies a relatively few key factors to which a security's return is assumed to be linearly related, in the following manner:

$$\bar{R}_i = a_i + \sum_l b_{il} \bar{F}_l + \bar{\varepsilon}_i \qquad (11)$$

In such a model the $\tilde{\varepsilon}_i$ values are assumed to be uncorrelated across securities. ROSS' (1976) Arbitrage Pricing Theory (APT) concludes that if returns are generated by such a model, expected returns must be approximately linearly related to the b_{il} values if opportunities to gain through arbitrage are to be precluded. However, the APT provides no implications concerning either the signs or the magnitudes of the coefficients in the associated pricing relationship.

It is entirely possible to augment the assumptions of the APT with those of the CAPM (most importantly, the assumption that investors maximize mean-variance utility functions). The resulting implications will then be consistent with both theories. Moreover, by making assumptions about investors' objectives one can obtain precise statements about the signs and magnitudes of the coefficients of the APT pricing relationship, as I have shown in SHARPE (1984).

The efficiency of the market portfolio

A key concept due to MARKOWITZ (1952) is that of the *efficiency* of a portfolio. In the present context a portfolio can be said to be *efficient* if it would be optimal for an investor with some (non-negative) risk tolerance. Comparison of equations (2) and (3) directly implies that the market portfolio is efficient in this sense.

Consider an investor who has a risk tolerance equal to τ_m and holds the market portfolio. Equation (3) shows that the first-order conditions for the maximization of his or her utility will be met for every security. Since the market portfolio will be optimal for such an investor, it must be efficient. More specifically, the market portfolio will be optimal for an investor with the average (societal) risk tolerance.

The Two-Fund Separation Theorem

Under the conditions assumed in the CAPM, every investor's optimal portfolio can be obtained by a suitably chosen combination of any two arbitrarily selected efficient portfolios. [17] Two natural choices are those that would be optimal for investors with risk tolerances of zero and τ_m. The former is the minimum-variance portfolio. The latter is, of course, the market portfolio. Following TOBIN (1969), this is generally termed the Two-Fund Separation Theorem.

When a riskless asset is available, the minimum-variance portfolio will be composed solely of that asset. Thus all investors will hold combinations of the riskless asset and the market portfolio. For investors with risk tolerance greater than τ_m, optimal investment will involve a negative position in the riskless asset and a positive position in the market portfolio - and hence all risky assets. Note, however, that the market portfolio will include the net positive supply of riskless assets in the economy; hence only investors with τ_k values considerably greater than τ_m will actually have to borrow money. Every other investor will select some combination of the riskless asset and the market portfolio, and thus require only non-negative holdings.

Key implication

The key implications of the CAPM are that:
1. the market portfolio will be efficient,
2. all efficient portfolios will be equivalent to investment in the market portfolio plus, possibly, lending or borrowing, and
3. there will be a linear relationship between expected return and beta.

Practical applications of these relationships are many. Investors can easily identify efficient portfolio strategies and such strategies can be effectively implemented through mutual fund and other institutional vehicles. Corporate and governmental decision-makers can use the Security Market Line relationship to determine the desirability of an investment project by comparing its expected return with that available in the capital market for projects with similar beta values (i.e., with similar market risk or sensitivity to economic conditions).

In the frictionless world of the CAPM, each investor chooses a portfolio that maximizes his or her utility.

This leads to an efficient distribution of risk in the economy, given, of course, the distribution of wealth among investors.

Negative Holdings

The CAPM assumes that investors can take negative positions in assets. For the riskless asset, the traditional manner in which such a position is achieved involves borrowing money. For a risky asset, the traditional method requires a *short sale*.

A "short position" is achieved by borrowing an asset such as a share of stock, with a promise to repay in kind, typically on demand. The borrowed asset is then sold, generating a cash receipt. If the proceeds of the sale may be used for other types of investment, the overall effect is equivalent to that of a negative holding of the asset in question. If, however, the proceeds are "impounded" to serve collateral for the borrowed asset, such a short position may differ from a negative holding of the asset in question. In many countries, proceeds from some short sales are impounded in this manner, and the short seller receives little or no interest on the impounded amount. [18] Moreover, some institutional investors are precluded from the use of short positions, either through explicit rules or implicit threat of suit for violation of fiduciary standards. Other restrictions may apply - for example, exchanges may not allow a short sale following a previous decline in the price of the security in question.

Of necessity, arbitrage-based theories allow short positions. Whether such positions will be taken in equilibrium is usually unclear, since the models lack sufficient assumptions (i.e., those concerning utility functions) to characterize equilibrium holdings. More importantly, by the very nature of the arbitrage approach, multiple solutions in terms of holdings are possible in equilibrium due to the presence (or potential presence) of redundant securities.

While the CAPM assumes that investors can take negative positions in any asset, it implies that in equilibrium the only such positions taken will involve the minimum-variance portfolio. When a riskless asset is available, the only negative holdings in equilibrium will involve borrowing by investors with above-average risk tolerance who wish to finance added investment in a portfolio representing the overall capital market.

With costless monitoring of investors' positions, zero transactions costs, and equal information about securities available to all investors, the only restriction on negative holdings would be those required to achieve payment of the requisite cash flows. An investor's overall portfolio must be such that in each state of the world the sum of the positive cash flows is at least as large as the sum of the negative cash flows required to match the payments made by the issuers of the securities held in negative amounts. If this condition is violated, such securities have not been fully replicated and the investor cannot expect to receive the full price when taking the associated negative positions.

Since all information is not fully public, and monitoring and transactions are costly, institutional arrangements for short selling have traditionally required the posting of separate collateral for each position, with little if any consideration given to the effects of diversification at the portfolio level. Under such conditions it is difficult or expensive for a high risk-tolerance investor to borrow money to finance added investment in a market-like portfolio. Stock index futures contracts now provide investors with a more efficient means for doing so. Before considering them, however, we investigate the characteristics of an extreme case in which no negative positions are allowed.

Capital Asset Prices without Negative Holdings

Assumptions
To explore the effects of constraints on asset holdings we retain all the assumptions of the CAPM and add $N \cdot K$ non-negativity constraints of the form:

$$X_{ik} \geq 0 \qquad \text{for all } i \text{ and } k.$$

Portfolio optimality

Investor k seeks to maximize U_k subject to a full investment constraint and the relevant non-negativity constraints. This is a *quadratic programming problem*. An exact solution to a portfolio optimization problem of this form can be obtained using the *critical line algorithm* developed by MARKOWITZ (1956). When the solution is obtained, some values of X_{ik} will be positive. The corresponding securities are said to be in the optimal portfolio. The remainder will be at their lower bounds of zero and are said to be *out* of the portfolio.

Each of the securities *in* the optimal portfolio must have the same marginal utility. If this were not the case, it would be possible to reallocate funds from one such security with a lower marginal utility to one with a higher marginal utility, thereby increasing utility without violating either the full investment constraint or any of the non-negativity constraints. The common value of marginal utility for such securities will be the investor's marginal utility of wealth, which we will again denote λ_{fk}.

Each of the securities *out* of the portfolio must have a marginal less than (or equal to) that of the securities in the portfolio. If this were not the case, it would be possible to reallocate funds from a security in the portfolio to one out of the portfolio, thereby increasing utility without violating any of the constraints. These relationships, which derive from more general ones termed *Kuhn-Tucker conditions*, can be written conveniently as:

$$E_i - \frac{2}{\tau_k} C_{ik} = \lambda_{jk} - z_{ik} \qquad \text{for all } i. \qquad (12)$$

where z_{ik} will be zero for securities in the portfolio and greater than or equal to zero for securities out of the portfolio.

Aggregation

As with the CAPM, we take a wealth-weighted average of the conditions for the optimality of individual investors' portfolios. The result has a similar from:

$$E_i - \frac{2}{\tau_m} C_{im} = \lambda_{fm} - z_{im} \qquad \text{for all } i. \qquad (13)$$

Not surprisingly, z_{im} is a weighted average of the z_{ik} values for the K investors, with the weights depending on the investors' influences in the market place, where influence depends on both wealth and risk tolerance. [19]

Expected returns

The previous equation can be transformed simply to obtain:

$$E_i = \lambda_{fm} + \frac{2V_m}{\tau_m} \beta_{im} - z_{im} \qquad \text{for all } i. \qquad (14)$$

Were it not for the last term, there would be a linear relationship between expected return and beta, as in the CAPM. But the last term implies that only securities that are in *every* investor's optimal portfolio will plot along such a line. [20] Every security that is *out* of at least one investor's optimal portfolio will plot below the line. [21] Moreover, the larger the number of investors for whom the lower bound for a security is binding, the larger is the corresponding value of z_{im} likely to be.

It is important to note that the magnitudes of the z_{im} values will be affected by the distribution of risk tolerances across investors. In the special case in which all investors have the same risk tolerance, everyone will choose to hold the market portfolio and all the z_{im} values will equal zero, giving results identical to those of the original CAPM. Loosely speaking, the greater the variation in risk tolerances across investors, the more likely it is that some of the z_{im} values will be positive. Note, however, that in this connection each investor's influence will depend on both his or her wealth and risk tolerance. Unless there is substantial variation in the risk tolerances of the wealthiest investors, the z_{im} values may be very close to zero, giving results very similar to those of the original CAPM.

The risk premium

To determine the risk premium in this case, one must take a market value-weighted average of the equilibrium relationships in the previous equation. Doing so gives:

$$\frac{E_m - \lambda_{fm}}{V_m} = \frac{2}{\tau_m} - \frac{z_{mm}}{V_m}, \tag{15}$$

where z_{mm} represents the market value weighted average of the z_{im} values. [22]

Clearly, the risk premium per unit of variance will be a function of the extent to which investors' portfolios are affected by the non-negativity constraints. [23]

The Security Market Line relationship

Given the final term in the previous equation, it is not particularly instructive to drive a counterpart to the Security Market Line of the CAPM. However, equation (14) can stand as the analogue to a Security Market Line for this case. As indicated earlier, the first two terms do indeed provide such a line. However, in this case the line serves as an upper boundary. Some or all of the securities may plot below the line, with the distances dependent on the degree to which the associated non-negativity constraints are binding. Thus there may not be a precise linear relationship between expected returns and beta values.

Riskless securities

In this case no riskless borrowing is allowed, since negative positions are precluded. The net positive supply of riskless securities will, however, be included in the market portfolio. For any investor for whom the constraint on borrowing is binding, λ_{fk} will exceed R_f. For all others, the two values will be equal. Hence, λ_{fm} will equal or exceed R_f.

The efficiency of the market portfolio

In the CAPM, the marginal utilities of all securities with respect to the holdings in the market portfolio are equal when evaluated using the societal risk tolerance. This is sufficient for the efficiency of the market portfolio. It is also necessary if the market portfolio is to be efficient for an investor with the societal risk tolerance.

More generally, for the market portfolio to be efficient, the marginal utilities of all securities measured relative to the market portfolio must be equal when evaluated using *some* positive risk tolerance. This follows from the fact that the market portfolio includes positive amounts of every security, hence all securities will be *in* such a portfolio.

For this condition to be met, there must be a linear relationship between security expected returns and their beta values. [24] With restrictions on negative holdings this may not be the case, due to the influences of the z_{im} values. [25] Hence the market portfolio may be inefficient. Of course, the extent of the inefficiency will depend on the magnitudes of the z_{im} values - if most are small, the degree of inefficiency of the market portfolio may be inconsequential.

Fund separation

MARKOWITZ (1959) showed that non-negativity constraints cause the efficient frontier to be piecewise linear in the space of holdings. Within each linear range, efficient portfolios can be obtained by combining any two other portfolios within the range. [26] However, no two portfolios can, in general, be utilized to obtain *all* efficient portfolios. Hence two-fund separation may not strictly apply in this case.

Key implications

When negative positions are precluded:
1. the market portfolio may not be efficient,
2. some efficient portfolios may not be equivalent to investment in the market portfolio plus, possibly, lending or borrowing, and
3. there may not be a linear relationship between expected return and beta.

All these implications suggest a diminution in the efficiency with which risk can be allocated in an economy. The choice of optimal portfolios becomes more difficult than in the simple setting of the CAPM. Calculations of cost of capital for corporate and governmental investment projects may require more than the determination of a simple relationship between expected return and market risk. More fundamentally, overall welfare may be lower than it would be if the constraints on negative holdings could be reduced or removed.

While the magnitudes of the departures from the implications of the original CAPM might be small even under the extreme conditions assumed in this case, it is clear that institutional arrangements to improve investor's abilities to take negative positions can increase the efficiency with which risk is allocated in an economy. Following some comments on financial innovation in general, I will discuss the stock index futures contract - an innovation that provides such an improvement.

Financial Innovation

More than most sciences, economics not only analyzes reality, it also alters it. Theory leads to empiricism which changes behavior. Nowhere is this more evident than in financial economics. The academic field of finance differs radically from that of three decades ago, due in large part to advances in financial economic theory and to the extensive empirical research that has flowed from those advances. At least as important, the practice of finance has been affected in fundamental ways by the progress in financial economics. Most notably, the last decade has been marked by unprecedented innovation in financial instruments, markets and institutions.

Given the bewildering pace of such innovation, it is not surprising that some individuals and organizations have at times found it difficult to fully understand the proper uses of some of the new instruments and procedures. Evidence abounds that those who fail to learn the principles of financial economics in more formal ways will do so through experience. Markets are effective although sometimes cruel teachers. In general, financial systems are self-correcting. Given time, participants learn to use new instruments and procedures to improve overall welfare, not just to reallocate wealth from one set of hands to another. It is usually best to wait until the forces of competition are able to regulate a market rather than to impose regulations prematurely.

Much financial innovation has been possible due to the remarkable advances in computation and com-

munication technology. Moreover, increased global competition, with the accompanying diminution of monopoly power on the part of organizations and governments, has played an important role. Nonetheless, I cannot help but believe that discoveries in the science of financial economics have had a major influence. Stock index futures contracts appear to provide a clear example of this.

Stock Index Futures Contracts

Features

A traditional *forward contract* is an agreement on the part of the seller to deliver a stated amount of a commodity on a given future date to the buyer at a pre-specified price. A *futures contract* is a standardized forward contract with the further provision that the delivery price be reset at the end of each trading day to equal the price at which new agreements were struck. At the time of each such resetting, one of the two parties pays the other an amount equal to the difference between the new price and the old. This process is known as "posting variation margin" as a result of the "marking to market" of the price of the contract.

The seller of a futures contract is said to be *short* the contract; the buyer is said to be *long*.

Futures exchanges make it possible for the two parties in a contract to be "unlinked". Thus A may sell a contract to B. B may later sell it to C without A's involvement. And later, C may sell it to A, extinguishing it before the final delivery date.

To protect the other party in such an arrangement, each party must post collateral as "margin". The amount need only cover potential losses between two trading days, however, due to the process of marking to market. Margins equal to 10% to 20% of the value of a position typically suffice. To insure that collateral is maintained, yet preserve the standardization of the contract, brokers representing those with positions utilize a *clearing house*, which provides assurance that obligations will be met.

A *financial futures* contract may call for the actual delivery of a stated financial instrument of one or

more of a set of such instruments. Alternatively, it may call only for a final marking to market on the delivery day, with one party paying the other an amount equal to the difference between the prior futures price and the value of the underlying financial instrument on the delivery day. The latter is often termed *cash delivery*.

A *stock index futures contract* covers a pre-specified portfolio of stocks. It allows investors to take long or short positions in diversified portfolios. Most such contracts provide for *cash delivery*.

Effects

A key aspect of a stock index futures contract is its focus on a *diversified portfolio* rather than on an individual security or commodity. In this respect, such a contract is similar to a mutual fund, unit trust, commingled index fund or stock index option. All provide investors with "packages" of securities, substantially reducing the costs associated with diversification. Recent growth in investors' reliance on all such vehicles stands, at least in part, as testimony to the influence of financial economics on the process by which risk is borne in modern economies. A second feature of a stock index futures contract is of particular interest in the present context. Such a contract provides an efficient method for simultaneously taking a positive position in a diversified stock portfolio and a negative position in a riskless asset. In effect, the purchaser of such a contract borrows money to purchase a stock portfolio while the seller lends money and takes a short position in the stock portfolio. [27] If the buyer of the contract posts as margin riskless securities with a value equal to that of the futures position, the net effect is similar to that of investing a comparable amount in the stocks in the associated index. If less than 100% margin is posted, the effect is similar to that of a levered purchase of the stocks in the index. While not precisely the same as borrowing at the riskless rate to purchase a portfolio of stocks, a long position in a futures contract can provide a very close approximation to such a strategy. Moreover, the upper limit on the borrowing implicit in the arrangement is sufficiently large to satisfy all but a potentially few

investors with extremely high risk tolerance. It is striking that a levered holding of a highly diversified portfolio (i.e., the market portfolio) is precisely the optimal investment strategy for high risk tolerance investors in the simple setting of the original CAPM. While no financial futures contract corresponds to the overall market portfolio, combinations of existing contracts (including those on bonds and other types of securities) may approximate such a strategy.

In a sense, the seller of a futures contract holds negative positions in the underlying securities, and the existence of such positions is inconsistent with the implications of the CAPM. However, when viewed more broadly, this inconsistency may be more apparent than real. Often the seller of a stock index futures contract also holds the individual securities that make up the index in question. The futures position thus provides a *hedge* against changes in the value of the portfolio of actual stocks. As a result, the hedged futures seller's *net* position is virtually risk-free and hence equivalent to investment in a riskless asset. Such a person can assemble the securities in an economical manner and, in effect, provide a means for others to buy or sell the package without incurring the costs associated with the purchase or sale of large numbers of individual stocks. In effect, he or she loans money to the high risk-tolerance investor to enable the latter to purchase added amounts of risky securities.

Futures contracts written on diversified portfolios require less margin than would a set of contracts on individual securities, each with its own required margin. Moreover, stock index contracts take advantage of economies of scale in transactions and record-keeping.

In effect, stock index futures contracts provide those who might be limited by traditional constraints on borrowing with a means for achieving desirable investment strategies. Moreover, the strategies are similar to those that are optimal for high risk-tolerance investors in the setting of the original CAPM. Hence, the existence of such contracts may well bring actual capital markets closer to those of this simple theory. If so, stock index futures con-

tracts may significantly improve the efficiency with which risk is allocated in an economy.

Conclusions

Here I have explicitly considered only one type of impediment to the efficient allocation of risk. However, this case can serve as representative of many others. The greater the costs and constraints associated with the purchase and sale of securities, the farther will an economy be from the goal of allocating risk to those most able and willing to bear it.

Happily, technological advances and greater understanding of the principles of financial economics are reducing costs and constraints of this type at a rapid pace. As a result, capital markets are moving closer to the conditions assumed in some of the simpler types of financial theory. Far more important: the combined efforts of theoreticians, empiricists and practitioners are increasing the efficiency with which risk is allocated among individuals, leading to improvements in social welfare.

Footnotes:

[1]   Beginning with SHARPE (1961) and SHARPE (1964).
[2]   More specifically, footnote [22].
[3]   Prominent arbitrage-based theories in financial economics are the BLACK/SCHOLES (1973) model, which deals with the prices of options vis-a-vis related securities, and ROSS' (1976) Arbitrage Pricing Theory, which draws implications about the prices of capital assets when returns are generated by a specified factor model. While the monumental ARROW (1953)-DEBREU (1959) state-preference approach to uncertainty makes some assumptions about individuals' utility functions, many of its key results are arbitrage-based. A number of theories in financial economics have been constructed using the Arrow-Debreu "state of the world" paradigm, among them the binomial model of option pricing first presented in my textbook SHARPE (1978) and then extended by COX, ROSS and RUBINSTEIN (1979) and many others.
[4]   It is, of course, true that arbitrage-based models make some assumptions about investor preferences - for example,

that investors prefer larger payoffs to smaller payoffs in any given state of the world. However, such assumptions are minimal, compared with the more detailed assumptions of utility-based models.

[5]   This is not required. The results that follow can be obtained under more general conditions, with the resultant values of τ_k interpreted as investors' marginal rates of substitution, given their optimal holdings.
[6]   Note, however, that this assumption will be less onerous, the shorter the time period under consideration. In continuous-time versions of the model, the time period is (in effect) infinitesimal in length. Under such conditions, two moments may serve as adequate representations even if the probability distributions of returns over a finite period and/or investors' preferences for returns over such a period are quite complex.
[7]   Aided, perhaps by an advisor.
[8]   The expected return of investors k's portfolio is given by:

$$E_k = \sum_i X_{ik} E_i$$

where X_{ik} represents the proportion of investor k's portfolio invested in asset i.

[9]   The variance of investor k's portfolio is given by:

$$V_k = \sum_i \sum_j X_{ik} X_{jk} C_{ij}$$

where C_{ij} represents the covariance between the returns on assets i and j.

[10]   Incorporating U_k and the full investment constraint in a Lagrangean function to be maximized gives:

$$Z_k = U_k + \lambda_{jk}\left(1 - \sum_i X_{ik}\right)$$

Clearly λ_{jk} is the *marginal utility of wealth* for investor k, since it equals the partial derivative of Z_k with respect to the investors' wealth (in this metric, the value 1 in the parenthesized full investment constraint, since the X_{ik} values are expressed as proportions of the investor's total wealth).

Recall that:

$$E_k = \sum_k X_{ik} E_i, \qquad \text{and}$$

$$V_k = \sum_i \sum_j X_{ik} X_{jk} C_{ij}$$

Hence, taking the partial derivative of Z_k with respect to investor k's holding of security i gives:

$$\frac{\partial Z_k}{\partial X_{ik}} = E_i - \frac{2}{\tau_k} \sum_j X_{jk} C_{ij} - \lambda_{jk}$$

But note that the covariance of security i with any portfolio p will equal a weighted average of the covariances of the

security with the securities in the portfolio, using relative portfolio holdings as weights. Here:

$$C_{ik} = \sum_j X_{jk} C_{ij}$$

Substituting this relationship in the prior equation and rearranging terms gives equation (2).

[11] To derive equation (3) begin by multiplying all terms in equation (2) by τ_k and rearranging slightly, giving:

$$\tau_k E_i - 2 C_{ik} = \tau_k \lambda_{jk}$$

Next, multiply all terms by W_k , then sum over all investors, giving:

$$\sum_k W_k \tau_k E_i - 2 \sum_k W_k C_{ik} = \sum_k W_k \tau_k \lambda_{jk}$$

Define τ_m as:

$$\tau_m \equiv \sum_k W_k \tau_k$$

Now consider the second term. Note that:

$$C_{ik} = Cov(\tilde{R}_i, \tilde{R}_k)$$

where $\tilde{R}_i$ and $\tilde{R}_k$ are, respectively, the returns on security i and investor k's portfolio. By the properties of covariance:

$$\sum_k W_k C_{ik} = Cov\left(\tilde{R}_i, \sum_k W_k \tilde{R}_k\right)$$

But the summation on the right-hand side is simply a wealth-weighted average of the returns on the investors' portfolios or, more simply put, the return on the *market portfolio*. Hence the summation on the left-hand side is the covariance of the return on security i with that of the market portfolio, which can be denoted C_{im}.

Making these substitutions and dividing all terms by τ_m gives:

$$E_i - \frac{2}{\tau_m} C_{im} = \frac{\sum_k W_k \tau_k \lambda_{jk}}{\tau_m}$$

The last term is a weighted average of the values of λ_{jk} for the K investors, with the weights given by the product of W_k and τ_k. Denoting this λ_{fm} gives equation (3).

[12] But not necessarily independent of.

[13] In fact, this cannot strictly be the case except in economies with infinitely many securities. Since the market-value weighted sum of the left-hand sides of equation (9) equals $\tilde{R}_m$, the market-weighted sum of the $\tilde{\varepsilon}_i$ values must be zero. Thus at least two of the $\tilde{\varepsilon}_i$ values must be negatively correlated.

[14] Taking expectations of equation (9) gives:

$$E_i = a_i + \beta_{im} E_m$$

Comparison with (8) implies that:

$$a_i = (1 - \beta_{im}) R_f$$

[15] Such as that suggested by MARKOWITZ (1959) that I further developed in SHARPE (1961, 1963 and 1970).

[16] In such a model the "common factor" can be highly correlated with the return on the market portfolio, but not precisely equal to it if the assumption that the residual values are to be uncorrelated with one another is to be maintained.

[17] To provide two-fund separation in this case, rewrite equation (2) as follows:

$$\sum_j 2 C_{ij} X_{jk} + \tau_k \lambda_{jk} = \tau_k E_i$$

Portfolio optimality requires that this relationship be satisfied for each of the N securities and that the full-investment constraint be met. This gives rise to a set of $N+1$ simultaneous equations that can be written as:

$$\begin{pmatrix} 2C_{11} & \cdots & \cdots & 2C_{1N} & 1 \\ . & & & & . \\ . & & & & . \\ 2C_{N1} & \cdots & \cdots & 2C_{NN} & 1 \\ 1 & \cdots & \cdots & 1 & 0 \end{pmatrix} \begin{pmatrix} X_{1k} \\ . \\ . \\ X_{NK} \\ \lambda_{jk}^* \end{pmatrix} = \begin{pmatrix} 0 \\ . \\ . \\ 0 \\ 1 \end{pmatrix} + \begin{pmatrix} E_1 \\ . \\ . \\ E_N \\ 0 \end{pmatrix} \tau_k$$

where λ_{jk}^* represents $\tau_k \lambda_{jk}$. Writing this in matrix notation gives:

$$DY = K + \tau_k F$$

The solution, obtained by multiplying both sides by the inverse of D is thus:

$$Y = \{D^{-1} K\} + \tau_k \{D^{-1} F\}$$

Note that the optimal portfolio is a linear function of τ_k. Thus it can be restated as a linear function of any two vectors satisfying the above equation associated with different values of τ_k.

[18] Often additional collateral must be "posted" as "margin", but the short seller is generally allowed to receive the proceeds associated with the investment of this amount.

[19] That is,

$$z_{im} \equiv \frac{\sum_k W_k \tau_k}{\tau_m} z_{ik}$$

[20] Only for such securities will all the z_{ik} values be zero, giving value of zero for z_{im} (their weighted average).

[21] At least one z_{ik} value will be positive; since the remainder will all equal zero, the value of z_{im} (their weighted average) will be positive.

[22] Multiplying equation (13) by X_{im} and summing over i gives:

$$E_m = \lambda_{fm} + \frac{2V_m}{\tau_m}\beta_{mm} - \sum_i X_{im} z_{im}$$

Rearranging terms gives equation (15).

[23] Moreover, the magnitude of λ_{fm} will be also be affected by such restrictions.

[24] For all securities in the market portfolio to have the same marginal utilities requires that:

$$E_i - \frac{2}{\tau}C_{im} = \lambda$$

where τ and λ are positive constants.

Rearranging terms and converting C_{im} to the equivalent value expressed in terms of β_{im} gives:

$$E_i = \lambda + \frac{2V_m}{\tau}\beta_{im}$$

To meet the condition for efficiency of the market portfolio, there must thus be a strictly linear relationship between E_i and β_{im}. Moreover, the intercept and slope must both be positive, otherwise the implied values of λ and τ, respectively will be negative.

[25] If all the z_{im} values are zero, of course, the linear relationship will obtain. In the highly unlikely event that there is a linear relationship between the z_{im} values and the corresponding beta values, there will also be a linear relationship between the E_i and β_{im} values.

[26] Those lying at the end-points provide convenient choices.

[27] For details see, for example, DUFFIE (1989).

References

ARROW, K. (1953): "Le Role des Valeurs Boursiers pour la Repartition la Meilleure des Risques", Econométrie, Colloques Internationaux du Centre National de la Recherche Scientifique, Vol. XI, Paris, pp. 41-47.

BLACK, F. (1972): "Capital Market Equilibrium with Restricted Borrowing", Journal of Business 45, pp. 444-454.

BLACK, F. and M. SCHOLES (1973): "The Pricing of Options and Corporate Liabilities", Journal of Political Economy 81, pp. 637-654.

BREEDEN, D. (1979): "An Intertemporal Asset Pricing Model with Stochastic Consumption and Investment Opportunities", Journal of Financial Economics 7, pp. 265-296.

BRENNAN, M.J. (1970): "Taxes, Market Valuation and Corporate Financial Policy", National Tax Journal, December, pp. 417-427.

COX, J., S. ROSS and M. RUBINSTEIN (1979): "Option Pricing: A Simplified Approach", Journal of Financial Economics 7, pp. 229-263.

DEBREU, G. (1959): "Theory of Value", John Wiley and Sons, Somerset, New Jersey.

DUFFIE, D. (1989): "Futures Markets", Prentice Hall, Englewood Cliffs, New Jersey.

GLENN, D. (1976): "Super Premium Security Prices and Optimal Corporate Financing Decisions", Journal of Finance 31, pp. 507-524.

KRAUS, A. and R. LITZENBERGER (1976): "Skewness Preference and the Valuation of Risk Assets", Journal of Finance, pp. 1085-1100.

LEVY, H. (1978): "Equilibrium in an Imperfect Market: A constraint on the Number of Securities in a Portfolio", American Economic Review 68, pp. 643-658.

LEVY, H. and H. MARKOWITZ (1979): "Approximating Expected Utility by a Function of Mean and Variance", American Economic Review, pp. 308-317.

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

LINTNER, J. (1969): "The Aggregation of Investors' Diverse Judgements and Preferences in Purely Competitive Markets", Journal of Financial and Quantitative Analysis 4, pp. 346-382.

MARKOWITZ, H. (1952): "Portfolio Selection", Journal of Finance 7, pp. 77-91.

MARKOWITZ, H. (1956): "The Optimization of a Quadratic Function Subject to Linear Constraints", Naval Research Logistics Quarterly 3, pp. 111-133.

MARKOWITZ, H. (1959): "Portfolio Selection: Efficient Diversification of Investments", John Wiley and Sons, Somerset, New Jersey.

MARKOWITZ, H. (1987): "Mean-Variance Analysis in Portfolio Choice and Capital Markets", Basil Blackwell Inc., New York.

MARKOWITZ, H. (1990): "Risk Adjustment", Journal of Accounting, Auditing and Finance, forthcoming.

MERTON, R. (1973): "An Intertemporal Capital Asset Pricing Model", Econometrica 41, pp. 867-887.

MERTON, R. (1987): "A Simple Model of Capital Market Equilibrium with Incomplete Information", Journal of Finance 52, pp. 483-510.

MODIGLIANI, F. and M. MILLER (1958): "The Cost of Capital, Corporation Finance, and the Theory of Investment", American Economic Review 48, pp. 261-297.

MOSSIN, J. (1966): "Equilibrium in a Capital Asset Market", Econometrica 35, pp. 768-783.

ROSS, S. (1976): "Arbitrage Theory of Capital Asset Pricing", Journal of Economic Theory 13, pp. 341-360.

RUBINSTEIN, M. (1974): "An Aggregation Theorem for

Securities Markets", Journal of Financial Economics 1, pp. 225-244.

SHARPE, W. (1961): "Portfolio Analysis Based on a Simplified Model of the Relationships Among Securities", PhD Disserta-tion, University of California at Los Angeles, June.

SHARPE, W. (1963): "A Simplifies Model for Portfolio Analysis", Management Science 9, pp. 277-293.

SHARPE, W. (1964): "Capital Asset Prices: A Theory of Market Equilibrium Under Conditions of Risk", Journal of Finance 19, pp. 425-442.

SHARPE, W. (1970): "Portfolio Theory and Capital Markets", McGraw-Hill, New York.

SHARPE, W. (1978): "Investments", Englewood Cliffs, Prentice-Hall, New Jersey.

SHARPE, W. (1984): "Factor Models, CAPMs, and the APT", Journal of Portfolio Management, Fall, pp. 21-25.

TOBIN, J. (1969): "Liquidity Preference as Behavior Towards Risk", Review of Economic Studies 25, pp. 65-86.

VON NEUMANN, J. and O. MORGENSTERN (1944): "Theory of Games and Economic Behavior", Princeton University Press, Princeton, New Jersey.