CFA Institute®
CFA Program

PORTFOLIO MANAGEMENT

CFA® Program Curriculum
2025 • LEVEL II • VOLUME 9

ISBN 978-1-961409-29-3 (paper)
ISBN 978-1-961409-40-8 (ebook)
August 2024

CONTENTS

How to Use the CFA Program Curriculum

The CFA® Program exams measure your mastery of the core knowledge, skills, and abilities required to succeed as an investment professional. These core competencies are the basis for the Candidate Body of Knowledge (CBOK™). The CBOK consists of four components:

- A broad outline that lists the major CFA Program topic areas (www.cfainstitute.org/programs/cfa/curriculum/cbok/cbok)
- Topic area weights that indicate the relative exam weightings of the top-level topic areas (www.cfainstitute.org/en/programs/cfa/curriculum)
- Learning outcome statements (LOS) that advise candidates about the specific knowledge, skills, and abilities they should acquire from curriculum content covering a topic area: LOS are provided at the beginning of each block of related content and the specific lesson that covers them. We encourage you to review the information about the LOS on our website (www.cfainstitute.org/programs/cfa/curriculum/study-sessions), including the descriptions of LOS "command words" on the candidate resources page at www.cfainstitute.org/-/media/documents/support/programs/cfa-and-cipm-los-command-words.ashx.
- The CFA Program curriculum that candidates receive access to upon exam registration

Therefore, the key to your success on the CFA exams is studying and understanding the CBOK. You can learn more about the CBOK on our website: www.cfainstitute.org/programs/cfa/curriculum/cbok.

The curriculum, including the practice questions, is the basis for all exam questions. The curriculum is selected or developed specifically to provide candidates with the knowledge, skills, and abilities reflected in the CBOK.

CFA INSTITUTE LEARNING ECOSYSTEM (LES)

Your exam registration fee includes access to the CFA Institute Learning Ecosystem (LES). This digital learning platform provides access, even offline, to all the curriculum content and practice questions. The LES is organized as a series of learning modules consisting of short online lessons and associated practice questions. This tool is your source for all study materials, including practice questions and mock exams. The LES is the primary method by which CFA Institute delivers your curriculum experience. Here, candidates will find additional practice questions to test their knowledge. Some questions in the LES provide a unique interactive experience.

DESIGNING YOUR PERSONAL STUDY PROGRAM

An orderly, systematic approach to exam preparation is critical. You should dedicate a consistent block of time every week to reading and studying. Review the LOS both before and after you study curriculum content to ensure you can demonstrate the knowledge, skills, and abilities described by the LOS and the assigned reading. Use the LOS as a self-check to track your progress and highlight areas of weakness for later review.

Successful candidates report an average of more than 300 hours preparing for each exam. Your preparation time will vary based on your prior education and experience, and you will likely spend more time on some topics than on others.

ERRATA

The curriculum development process is rigorous and involves multiple rounds of reviews by content experts. Despite our efforts to produce a curriculum that is free of errors, in some instances, we must make corrections. Curriculum errata are periodically updated and posted by exam level and test date on the Curriculum Errata webpage (www.cfainstitute.org/en/programs/submit-errata). If you believe you have found an error in the curriculum, you can submit your concerns through our curriculum errata reporting process found at the bottom of the Curriculum Errata webpage.

OTHER FEEDBACK

Please send any comments or suggestions to info@cfainstitute.org, and we will review your feedback thoughtfully.

Portfolio Management

LEARNING MODULE

1

Economics and Investment Markets

by Andrew Clare, PhD, and Thomas F. Cosimano, PhD.

Andrew Clare, PhD, is at Cass Business School (United Kingdom). Thomas F. Cosimano, PhD, is Professor Emeritus in Finance at the Mendoza College of Business, University of Notre Dame, and Visiting Scholar, International Monetary Fund (USA).

LEARNING OUTCOMES

Mastery	*The candidate should be able to:*
☐	explain the notion that to affect market values, economic factors must affect one or more of the following: 1) default-free interest rates across maturities, 2) the timing and/or magnitude of expected cash flows, and 3) risk premiums
☐	explain the role of expectations and changes in expectations in market valuation
☐	explain the relationship between the long-term growth rate of the economy, the volatility of the growth rate, and the average level of real short-term interest rates
☐	explain how the phase of the business cycle affects policy and short-term interest rates, the slope of the term structure of interest rates, and the relative performance of bonds of differing maturities
☐	describe the factors that affect yield spreads between non-inflation-adjusted and inflation-indexed bonds
☐	explain how the phase of the business cycle affects credit spreads and the performance of credit-sensitive fixed-income instruments
☐	explain how the characteristics of the markets for a company's products affect the company's credit quality
☐	explain the relationship between the consumption hedging properties of equity and the equity risk premium
☐	explain how the phase of the business cycle affects short-term and long-term earnings growth expectations
☐	describe cyclical effects on valuation multiples
☐	describe the economic factors affecting investment in commercial real estate

1 INTRODUCTION

The state of the economy and financial market activity are interconnected. Financial markets are the forums where savers are connected with investors. This activity enables savers to defer consumption today for consumption in the future, allows governments to raise the capital necessary to create a secure society, and permits corporations to access capital to exploit profitable investment opportunities, which, in turn, should help to generate future economic growth and employment. Furthermore, all financial instruments essentially represent claims on an underlying economy. There is, therefore, an important and fundamental connection that runs from the decisions of economic agents, as they plan their present and future consumption, to the prices of financial instruments, such as bonds and equities.

The purpose of this reading is to identify and explain the links between the real economy and financial markets and to show how economic analysis can be used to develop ways of valuing both individual financial market securities and aggregations of these securities, such as financial market indexes. We begin by reviewing what we refer to as the fundamental pricing equation for all financial instruments. Using this framework, we then move on to explore the relationship between the economy and real default-free debt. From there, we can extend the analysis to the ways in which the economy can influence the prices of the following: nominal default-free debt; credit risky debt (for example, corporate bonds); publicly traded equities; and commercial real estate.

2 THE PRESENT VALUE MODEL

☐ explain the notion that to affect market values, economic factors must affect one or more of the following: 1) default-free interest rates across maturities, 2) the timing and/or magnitude of expected cash flows, and 3) risk premiums

The reference point for the analysis of this reading is the present value model of asset valuation. The impact of economic factors on asset values can be studied in the context of that model by examining how economic factors can affect discount rates and future cash flows. These topics are explored in more detail in the following sections.

The Present Value Model

The value of an asset must be related to the benefits that we expect to receive from holding it; for many assets (e.g., financial securities), these benefits are its future cash flows, which may be specified in the security's contract, as is the case with bonds, or be discretionary, as is the case with ordinary shares. Intuitively, a given amount of money received in the future will be valued less by individual investors than the same amount of money received today. Because an investor can use cash for present consumption, he or she needs an incentive to defer it to the future and more so as the future becomes less certain. These considerations provide an economic rationale for valuing an asset by discounting its future cash flows to derive its present value.

Equation 1 presents the fundamental present value formula for the value at time t of any financial asset i, V_t^i, which we assume equals its current market price, P_t^i. In general, we will speak of the time t as "today."

$$P_t^i = \sum_{s=1}^{N} \frac{E_t\left[\widetilde{CF}_{t+s}^i\right]}{\left(1 + l_{t,s} + \theta_{t,s} + \rho_{t,s}^i\right)^s}, \quad (1)$$

where

P_t^i = the value of asset i at time t (today)

N = number of cash flows in the life of the asset

$\widetilde{CF}_{t+s}^i$ = the uncertain, nominal cash flow paid s periods in the future

$E_t[\widetilde{CF}]$ = the expectation of the random variable $\widetilde{CF}$ conditional on the information available to investors today (t)

$l_{t,s}$ = yield to maturity on a real default-free investment today (t), which pays one unit of currency s periods in the future

$\theta_{t,s}$ = expected inflation rate between t and $t + s$

$\rho_{t,s}^i$ = the risk premium required today (t) to pay the investor for taking on risk in the cash flow of asset i, s periods in the future

This expression is general enough to be used to value all financial instruments. The present values of all of the instrument's cash flows are summed from 1 to N. Some assets, such as a five-year zero-coupon bond, may have only one cash flow, and so N would equal five in that case, with cash flows in Periods 1–4 equal to zero. At the other extreme, dividend-paying equities produce cash flows in the form of dividends into the indefinite future, in which case N could, technically, be equal to infinity.

According to Equation 1, effects of the economy on asset prices are transmitted through some combination of influences on the numerator—the asset's expected cash flows—and denominator—the discount rate(s) applied to the asset's expected cash flows.

A factor that typically distinguishes one financial asset class from another is the degree of certainty that investors have about future cash flows. At one extreme there may be little uncertainty. For example, despite losing its AAA rating from Standard & Poor's, investors might still attach a relatively low probability to the prospect of the US Treasury not making the scheduled payments on its debts on time and in full. Investors may regard the probability of the German government defaulting on its debts to be very low, too. At the other extreme, investors may be very uncertain about the size and timing of dividend payments from an equity investment and will also have to consider the prospect of receiving no dividends at all in the event that the company declares bankruptcy.

The uncertainty about future cash flows is reflected in the discount rate in Equation 1. We can think of the discount rate as having three distinct components. The first component is $l_{t,s}$ which effectively represents the return that an investor requires on a real default-free fixed-income security at present time t for a cash flow to be paid s periods in the future. For example, readers can think of this return as being analogous to the return expected on an investment in an inflation-linked bond issued by the government of a developed economy.

The second component in the discount rate, $\theta_{t,s}$, represents the additional return required by investors, above that required from investing in a real default-free investment, for investing in a nominal default-free investment. This additional return is required even though an investor may attach a zero probability to not being paid on time and in full, because future nominal payments will be affected by inflation. In essence, this component of the discount rate represents the compensation that investors demand for the inflation that they expect to experience over the investment horizon. Compensation is demanded because investors are concerned about the real purchasing power of their investments in the future rather than in the nominal value of the future cash flows.

The third component of the discount rate in Equation 1, $\rho_{t,s}^{i}$, represents the additional return that investors expect for investing in financial assets because of uncertainty about the asset's future cash flows. In other words, all securities, even those issued by governments of developed economies and considered risk free in that there is negligible risk that the issuer will default (as we will see later), carry some risks for which risk-averse investors will want to be compensated. Indeed, the size of this risk premium will vary among asset classes, and this variation is largely responsible for the distinction between one asset class and another. We remind the reader of this difference among assets by placing a superscript i to indicate that this premium is specific to the asset under consideration.

As we will show throughout this reading, the size and nature of this addition to expected return, $\rho_{t,s}^{i}$, will depend on the characteristics of the asset or asset class in question, which, in turn, will be determined by developments and expected developments in the real economy. This means that the discount rates applied to the cash flows of financial assets will almost certainly vary over time as perceptions of expected economic growth, inflation, and cash flow risk change. In particular during recessions, the risk premium that investors demand on financial assets, especially those that are not default-free, may rise because investors in general may be less willing and able to take on heightened default risk during such periods.

The $\rho_{t,s}^{i}$ component may include more than just the compensation for the uncertainty related to financial cash flows that may be subject to default risk. In particular, $\dot{\rho}_{t,s}$ may also reflect other types of risk—for example, liquidity risk. Liquidity risk refers to the possibility that a financial asset cannot be converted quickly into cash at close to its fair value; it is particularly characteristic of investments in commercial real estate and high-yield corporate bonds/loans. And as many investors learned during the 2008–2009 global financial crisis, some debt instruments, such as mortgage-backed securities, can become very illiquid at just the moment when investors become most risk averse and when they want to be holding cash rather than riskier financial assets.

In summary, the expected cash flows for any financial asset, i, can be discounted using the following general expression for the discount rate (this additive expression is an approximation of the exact multiplicative expression that includes interaction effects between the terms):

$$1 + l_{t,s} + \theta_{t,s} + \rho_{t,s}^{i} \text{ for } s = 1, \quad \ldots, \quad N. \tag{2}$$

A major purpose of this reading is to identify the relationship between these elements of the discount rate and the underlying economy and also to decompose $\dot{\rho}_{t,s}$ into its component parts for each asset class.

3 EXPECTATIONS AND ASSET VALUES

☐ explain the role of expectations and changes in expectations in market valuation

Examining Equation 1 reveals simple but important observations:

- Asset values depend not on past cash flows but on the expectation of future cash flows.

- These expectations are based on (conditional on) current information (indicated by the time subscript *t*) that may be relevant to forecasting future cash flows. Any information that may contribute to the accuracy or precision of expectations is relevant.

Because asset values are dependent on expectations of future cash flows, information that changes expectations affects asset values and realized returns. Information that has been anticipated is already reflected in asset prices, but information that is different from what was expected constitutes real news that requires expectations to adjust. The adjustment generates a holding period return that differs from the expected return. This observation about investors' anticipations is important to understanding sometimes seemingly counterintuitive market reactions to economic information releases. Investors judge economic data releases relative to their expectations for the data. Prices may fall (rise) despite "good" ("bad") news if the expectation was for better (worse) news.

Thus, for valuation, one important distinction is information that is "news" or new information and information that has been fully anticipated. Therefore, news is a surprise relative to fully anticipated information.

Although this reading's focus is the effect of economic factors on asset values, investor sentiment (e.g., enthusiasm or despair) can also affect asset values. Economic factors affect asset values through generally direct effects on cash flows and/or discount rates, but investor sentiment affects asset values through direct effects on discount rates via higher or lower risk premiums (and possibly indirect effects on future cash flows).

4 THE DISCOUNT RATE ON REAL DEFAULT-FREE BONDS: INTEREST RATES

☐ explain the notion that to affect market values, economic factors must affect one or more of the following: 1) default-free interest rates across maturities, 2) the timing and/or magnitude of expected cash flows, and 3) risk premiums

Having introduced several fundamental concepts, we can now begin the analysis of economic factors affecting asset values in detail. The first step in understanding the relationship between the economy and investment markets involves noting that the purchase of an investment involves the opportunity cost of lower consumption today. In other words, by buying a financial asset, an investor defers some current consumption. The sum of all of these individual saving and consumption decisions is thus going to have an impact on the price of financial assets.

> *To explain how investors' concerns for satisfying consumption needs through economic fluctuations affect asset prices, modern finance makes reference to several expressions (*Equation 3*through*6*) that model how investors evaluate consumption trade-offs. These expressions, and associated (non-testable) calculations, show the analysis behind statements about the relationships between real interest rates, GDP growth, and the volatility of GDP; verbal statements of the intuition behind the relationships are also given that aid in understanding the entire reading.*

Real Default-Free Interest Rates

To demonstrate the importance of the aggregation of these individual saving and investment decisions, we can think through how the aggregated consumption and investment decisions of individuals might determine the real default-free interest rate in an economy—that is, $l_{t,s}$ in Equation 1. Consider a single individual who has to choose between using some portion of his or her wealth to consume today (t) or investing that wealth in default-free bonds that will pay investors one dollar when they mature s periods in the future.

Think of this bond as being issued by a highly rated, developed-economy government, so there is only a negligible prospect of default. Also think of this bond as being inflation index-linked so that investors do not need to concern themselves with the impact of future inflation on the bond's future cash flow. An alternative way of conceptualizing the bond is to assume (only for the moment) that it is issued by this government in a world that has no inflation. Later we will re-introduce inflation into the pricing problem.

What sort of return would investors require on a bond that is both default-free and unaffected by future inflation? It is tempting to say that an investor would require no return on such a bond because there is no risk of losing money over the investment period in either nominal or real terms. But the choice to invest today involves the opportunity cost of not consuming today. It is the aggregated opportunity cost of all investors that will determine the price of this asset today and its return over the investment horizon.

Think of the return on the asset as the opportunity cost (price) of consuming today. If the return increases, the investor substitutes away from current consumption to future consumption by purchasing an asset. Consequently, as with any other economic decision, an investor must consider the relative prices of the two alternatives. In this case, the investor can

- pay price $P_{t,s}$ today, t, of a default-free bond paying 1 monetary unit of income s periods in the future or
- buy goods worth $P_{t,s}$ dollars today.

The decision to purchase this bond will be determined by the willingness of individuals to substitute consumption today for consumption in the future. This trade-off is measured by the marginal utility of consumption s periods in the future relative to the marginal utility of consumption today (t). The marginal utility of consumption is the additional satisfaction or utility that a consumer derives from one additional unit of consumption. The ratio of these two marginal utilities—the ratio of the marginal utility of consumption s periods in the future (the numerator) to the marginal utility of consumption today (the denominator)—is known as the **inter-temporal rate of substitution**, denoted $\widetilde{m}_{t,s}$.

In "good" economic times, individuals may have relatively high levels of current income so that current consumption is high. In this case, the utility derived from an additional unit of consumption today will be relatively low. Conversely, in "bad" economic times, current income and consumption will tend to be relatively low, which means that the utility derived from an additional unit of consumption today will be relatively high. In addition, the marginal utility of consumption of investors diminishes as their wealth increases because they have already satisfied fundamental needs. Thus, investors would receive a larger benefit (utility) from an asset that pays off more in bad economic times relative to one that pays off in good economic times.

The rate of substitution is a random variable because an investor will not know how much she has available in the future from other sources of income, such as salary from working. This uncertainty is present even for an investment that pays a certain amount in one period because the value of this investment is determined by how much utility the investor receives from this investment.

Given this uncertainty, the investor must make the decision today based on her expectations of future circumstances when she receives the payoff from the investment. This expectation is conditional on the information that the investor has when the decision is made. Thus, if the investor wanted to consider an investment in a zero-coupon bond at time t that is certain to pay off one unit of real consumption in s periods, then

$$P_{t,s} = E_t\left(1\,\widetilde{m}_{t,s}\right) = E_t\left(\widetilde{m}_{t,s}\right) \quad (3)$$

where $\widetilde{m}_{t,s}$ is the investor's marginal willingness to trade consumption at time t for (real) wealth at time $t + s$.[1]

EXAMPLE 1

The Inter-Temporal Rate of Substitution (1 of 3)

1. Suppose the investor's willingness to trade present for future consumption can be represented as

$$\widetilde{m}_{t,1} = e^{a+b\tilde{z}}$$

Here, $\tilde{z}$ is a random shock to the economy that affects the cash flows of the marginal investor; $\tilde{z}$ is what makes $\widetilde{m}_{t,1}$ a random variable. The exponential form is consistent with assumptions about investor risk aversion and consumption growth often made in finance; a and b are typically negative given those same assumptions.[2] Parameters of the distribution of $\tilde{z}$ consistent with observed market data can also be established. Suppose that $\tilde{z}$, assumed to have a mean of zero, takes on one of two values—a negative value indicating a bad state or a positive value indicating a good state. The probabilities of bad and good states are 0.4 and 0.6, respectively.

Using market-consistent values in the exponent in the expression for $\widetilde{m}_{t,1}$, we can calculate the price of a bond promising $1 for sure in one year as the expected value of the investor's willingness to trade present for future consumption:

1 The term $\widetilde{m}_{t,s}$ is technically defined as $(\delta)[MU(C_{t+s})/MU(C_t)]$, where MU denotes marginal utility of consumption, C, and δ is a discount factor that captures the preference for consumption at t rather than later at t + s. The discount factor applied to MU (C_{t+s}) adjusts it for the time difference of s periods. The tilde on $\widetilde{m}_{t,s}$ indicates it is a random (stochastic) variable, and the term may be called the stochastic inter-temporal rate of substitution. As in Example 1, this inter-temporal rate of substitution varies based on the realization of the shock to economic activity, $\tilde{z}$. In the modern theory of asset pricing, the term m is also referred to as the stochastic discount factor or pricing kernel.

2 The negative exponential function expression reflects constant relative risk aversion utility and lognormally distributed consumption growth. In e^{a+bz}, the expected value is a and standard deviation is z. For an explanation of constant relative risk aversion, see an investments text such as Elton, Gruber, Brown, and Goetzmann (2014).

$$E_t\left(\widetilde{m}_{t,1}\right) = 0.4\,e^{a+b\times(z\text{ for a bad state})} + 0.6\,e^{a+b\times(z\text{ for a good state})}$$
$$= 0.4 \times 0.954676 + 0.6 \times 0.954379 = 0.954498$$

In the calculation, 0.954676 and 0.954379 are the asset's prices in the bad and good states, respectively. Note that the values for the random shock are consistent with the level of the yield curve in the United States from January 1999 to January 2014. The derivation of these numbers is beyond the scope of this reading. Also note the higher value of $1 received in the bad state. Following Equation 3, the investor is willing to buy the risk-free bond today for $0.954498 in exchange for $1 in one year. Also notice that the willingness to invest is smaller for the positive shock (z for a good state), because an investor is willing to pay less for the bond in the case of a good state. Thus, the positive shock is associated with a higher level of consumption today by the investor.[3]

The investor knows that she cannot affect the price of the bond, and so she must decide whether to buy or sell the bond based on this given price, $P_{t,1}$ (0.954498 from Example 1). If this price of the bond was less than the investor's expectation of the inter-temporal rate of substitution (suppose this is 0.9560), then she would prefer to buy more of the bond today. As more bonds are purchased, today's consumption falls and marginal utility of consumption today rises, so expectations conditional on current information of the inter-temporal rate of substitution, $E_t\left(\widetilde{m}_{t,s}\right)$, fall. This process continues until the rate of substitution is equal to the bond price shown in Equation 3; that is, equivalently, 0.9560 would fall and converge on 0.954498.

It is worthwhile to emphasize this point: All investors are essentially making investment decisions using Equation 3; some will want to sell their bonds to fund additional, current consumption, whereas others will want to buy bonds and defer some additional consumption until the future. To demonstrate the link between the bond price and these consumption/investment decisions, imagine for the moment that the market price of this bond is too "low" for an individual investor. In this case, the investor with a higher initial inter-temporal rate of substitution (higher $\widetilde{m}_{t,s}$) would buy more of the bond. As a result of this purchase, the investor will consume less today, leading to an increase in today's marginal utility, but he or she would expect to have more consumption and thus lower marginal utility in the future. Consequently, the inter-temporal rate of substitution would fall.

One investor cannot influence the equilibrium price. But if a substantial group of investors responded this way, then the demand and price of the bond would rise; in the illustration earlier, it would mean that it is possible for the price of the bond, 0.954498, to rise at the same time that the individual investor's inter-temporal rate of substitution was falling. This process would continue until all investors' willingness to invest converges on a single equilibrium value so that Equation 3 is true for all individuals and the market price is determined.

Conversely, if the market price of the bond were too "high" for a group of investors, then the investors with a lower inter-temporal rate of substitution would buy less of the bond. They would have more consumption and lower marginal utility today, but they would expect to have less consumption and higher marginal utility in the future. As a result, the inter-temporal rate of substitution would rise and the demand and price of the bond would fall. This process would again continue until Equation 3 is true for all individuals.

3 An exponential function, $f(x) = e^x$, is always increasing in the variable x and increases at an increasing rate as x gets large. For example, $f(0) = e^0 = 1$ and $f(0.05) = e^{0.05} = 1.0513$. If x is negative, the more negative x is, the smaller the value of the function. For example, compare $f(-0.05) = e^{-0.05} = 0.9512$ and $f(-0.02) = e^{-0.02} = 0.9802$; the function is still increasing in x.

EXAMPLE 2

The Inter-Temporal Rate of Substitution (2 of 3)

In Example 1, suppose the current market price of the real default-free bond is $9,540 per $10,000, but the investor's inter-temporal rate of substitution is $0.954498 per $1 promised. The investor would then value the guarantee of $10,000 in one period more than the market, so she would purchase it. As she buys more of the bond, her future income will be higher and its marginal utility lower, leading to a fall in her marginal willingness to invest in the risk-free asset. Only if there are many investors with the willingness to trade at $9,544.98 would the market price increase until all investors have the same marginal willingness to invest.

In summary, all investors use Equation 3 to make their investment decisions, so the equilibrium price in the market for these bonds equals the expectation of the inter-temporal rate of substitution of every single investor who participates in the bond market.

If the investment horizon for this bond is one year and the payoff then is $1, the return on this bond can be written as the future payoff minus the current payment relative to the current payment:[4]

$$l_{t,1} = \frac{1 - P_{t,1}}{P_{t,1}} = \frac{1}{E_t\left(\tilde{m}_{t,1}\right)} - 1 \quad (4)$$

Consequently, the return is higher for lower current prices. Equation 4 implies that the one-period real risk-free rate is inversely related to the inter-temporal rate of substitution. That is, the higher the return the investor can earn, the more important current consumption becomes relative to future consumption.

EXAMPLE 3

The Inter-Temporal Rate of Substitution (3 of 3)

Following the circumstances in Example 1, the one-period real risk-free interest rate is $l_{t,1} = \frac{1 - 0.954498}{0.954498} = 0.047671$, or 4.7671%.

5 THE DISCOUNT RATE ON REAL DEFAULT-FREE BONDS: UNCERTAINTY AND RISK PREMIUMS

☐ explain the notion that to affect market values, economic factors must affect one or more of the following: 1) default-free interest rates across maturities, 2) the timing and/or magnitude of expected cash flows, and 3) risk premiums

4 The step from the first expression to the second follows from rearranging the first expression, $\frac{1 - P_{t,1}}{P_{t,1}}$, as $\frac{1}{P_{t,1}} - 1$ and then substituting from Equation 3.

An investor's expected marginal utility associated with a given expected payoff is decreased by any increase in uncertainty of the payoff; thus, the investor must be compensated with a higher expected return. This result follows from decreasing marginal utility of wealth or income because the loss of utility from lower wealth is larger than the gain from an equivalent increase in wealth. The risk premium compensates the investor for the loss from this fluctuation in future wealth or income. An individual who requires compensation for this uncertainty is called "risk averse." This property was seen in Example 1, in which the inter-temporal rate of substitution was lower in the good state of the economy compared with the bad state.

For the valuation of cash flows under uncertainty, a second property of most investors' utility is important. In particular, an investor's absolute risk aversion is assumed to fall if he or she has higher wealth or income. Absolute risk aversion relates to the amount held in risky assets at different levels of wealth; under the assumption of decreasing absolute risk aversion made here, an investor invests larger amounts in risky assets as wealth or income increases (note that absolute risk aversion is in contrast to relative risk aversion, which relates to the fraction, not the amount, of wealth held in risky assets at different levels of wealth). Consequently, one's marginal utility is always lower as one's wealth or income increases. In this case, the risk premium for a given risk is lower for wealthier individuals because the average loss of marginal utility (slope of utility) from any risk taking is smaller, which means that relative to poorer individuals, wealthier individuals are more willing to take on a given risk. Consequently, wealthier investors are willing to buy more risky assets because they would value the asset more than poorer investors. But the expected marginal utility for wealthier investors will decline as they buy more of the risky asset. Eventually, both the wealthier and poorer investors would have the same willingness to invest in risky assets when the financial market is in equilibrium.

EXAMPLE 4

The Case of Increasing Wealth

1. This idea can be illustrated by raising the economic shock by a fixed amount regardless of whether the economy is good or bad, which has the effect of increasing the individual's resources and making her wealthier. For example, suppose we add 0.1 to $\tilde{z}$ and thus to the resources of the investor relative to the shock in Example 1. The expected inter-temporal rate of substitution for the investor is now lower for this safe asset (the default-free bond). The expected value of the investor's willingness to trade present for future consumption would then be

   $$E_t\left(\tilde{m}_{t,1}\right) = 0.4 \times 0.954528 + 0.6 \times 0.954231 = 0.954350.$$

   Compare this result with Example 1. The inter-temporal rate of substitution is lower under the good and bad shock to the economy. As a result, the expected inter-temporal rate of substitution $E_t\left(\tilde{m}_{t,1}\right)$ is lower for the wealthier investor by 0.000148 (= 0.954498 – 0.954350). Thus, the wealthier investor will buy the safe bond only at a lower price, and if this lower price is not the equilibrium price, the investor will substitute away from riskless assets to risky assets. Because of decreasing absolute risk aversion with wealth and because their fundamental consumption needs are met, wealthy investors will demand a lower premium than poorer investors for holding risky assets, all else being equal.

An individual with decreasing absolute risk aversion would lower the price of safe assets (see Altug and Labadie [2008] for a derivation of this result). If the rich individuals are a large percentage of the market, then the equilibrium return on the safe asset increases with the lower price. As a result, the poorer individuals would have incentives to increase their savings with the expected higher return on the safe asset. These savings allow all investors to partially compensate for any additional losses during possible bad times. Consequently, all the investors in the financial market would increase their savings when uncertainty about their future income increases. This higher savings means that the expected marginal utility in the future is lower because the investors' future resources are higher. Thus, the equilibrium price based on Equation 3 is lower, meaning that investors are compensated with a higher expected return when uncertainty in income increases.

6 THE DISCOUNT RATE ON REAL DEFAULT-FREE BONDS: RISK PREMIUMS ON RISKY ASSETS

☐ explain the notion that to affect market values, economic factors must affect one or more of the following: 1) default-free interest rates across maturities, 2) the timing and/or magnitude of expected cash flows, and 3) risk premiums

The price of other (non-default-free) financial instruments is established relative to the price of the default-free bond. This relationship can be seen by considering a default-free bond with a maturity of s periods (s is greater than or equal to two). Assume that the investor is holding the security for only one period. Its current price is $P_{t,s}$. In this case, the bond has value $\tilde{P}_{t+1,s-1}$ in one period because the term to maturity of the bond has been reduced by one period relative to its original maturity date. As a result, the investor's decision is now given by

$$P_{t,s} = E_t\left(\tilde{P}_{t+1,s-1}\tilde{m}_{t,1}\right) \quad (5)$$

The price in one period is uncertain because the s period bond is sold at the market price before it matures. Also notice that there is no interest payment because the bond promises a payment only at the terminal time. If a coupon is promised at time $t + 1$, then its value would have to be added to the right-hand side of Equation 5.

EXAMPLE 5

Pricing a Two-Period Default-Free Bond

1. In this example, we illustrate how the pricing formula in Equation 5 leads to a risk premium on a two-period default-free bond that is not present in the one-period default-free bond. In these calculations, we use five or six digits to the right of the decimal point because the risk premium is small for a two-period bond relative to a one-period default-free bond.

   Suppose the price at Time 1 of the two-year default-free bond is given by

$$\tilde{P}_{t+1,2-1} = e^{a'+b'\tilde{z}}$$

In this case, the future price can be shown to be

$\tilde{P}_{t+1,2-1} = 0.839181$ for \$1 at Time 2 with probability $p = 0.4$ and

$\tilde{P}_{t+1,2-1} = 0.954840$ for \$1 at Time 2 with probability $p = 0.6$.

The expected price at time $t + 1$ of a \$1 bond maturing at time $t + 2$ is 0.4 × 0.839181 + 0.6 × 0.954840 = \$0.908576. Without considering the investor's willingness to invest, the current value of the two-period bond is the simple present value using the one-period real risk-free interest rate of 4.7671% (from Example 3) as the discount rate. Thus, under the assumption stated, the bond would be worth $\frac{E_t(\tilde{P}_{t+1,s-1})}{1+l_{t,1}} = \frac{0.908576}{1.047671} = \$0.867234$. But the actual price in the financial markets based on Equation 5 is

$$P_{t,s} = E_t(\tilde{P}_{t+1,s-1}\tilde{m}_{t,1})$$

$$= 0.4 \times 0.839181 \times 0.954676 + 0.6 \times 0.954840 \times 0.954379$$

$$= 0.867226,$$

where 0.954676 and 0.954379 are the asset's prices in the bad and good states, as determined in Example 1. The price based on Equation 5 is smaller than the present discounted value at the risk-free rate; the difference is 0.000008 per 1 principal value (i.e., 0.867234 – 0.867226 = 0.000008). Thus, the holder of a two-year bond earns a risk premium. The reason for this result can be seen by calculating

$$E_t(\tilde{P}_{t+1,s-1})E_t(\tilde{m}_{t,1}) = 0.908576(0.4 \times 0.954676 + 0.6 \times 0.954379)$$

$$= 0.867234,$$

where 0.908576 is the Time 1 price of the bond as determined earlier. Consequently, we see that

$$E_t(\tilde{P}_{t+1,s-1})E_t(\tilde{m}_{t,1}) > E_t(\tilde{P}_{t+1,s-1}\tilde{m}_{t,1})$$

To summarize, the price uncertainty of the two-period bond at $t = 1$ gives rise to a risk premium, although the bond is default-risk free.

Example 5 showed how future price uncertainty creates a discount for risk. We now derive an alternative expression for the pricing relationship in Equation 5 that explains the nature of that discount and sheds further light on the conclusion of Example 5. In statistics texts, the following relationship between expected values and covariance is proven:

$$E_t(\tilde{x}\tilde{y}) = E_t(\tilde{x})E_t(\tilde{y}) + \text{cov}(\tilde{x},\tilde{y})$$

Here, $\text{cov}_t(\tilde{x},\tilde{y})$ refers to the conditional (on information at time t) covariance of the random variable $\tilde{x}$ with $\tilde{y}$. Thus, from Equation 5,

$$P_{t,s} = E_t(\tilde{P}_{t+1,s-1}\tilde{m}_{t,1}) = E_t(\tilde{P}_{t+1,s-1})E_t(\tilde{m}_{t,1}) + \text{cov}(\tilde{P}_{t+1,s-1},\tilde{m}_{t,1}).$$

But from Equation 4, $1 + l_{t,1} = \frac{1}{E_t(\widetilde{m}_{t,1})}$. So, an alternative way to view the pricing relationship in Equation 5 is

$$P_{t,s} = \frac{E_t(\widetilde{P}_{t+1,s-1})}{1 + l_{t,1}} + \text{cov}_t(\widetilde{P}_{t+1,s-1}, \widetilde{m}_{t,1}), \quad (6)$$

where $\text{cov}_t(\widetilde{P}_{t+1,s-1}, \widetilde{m}_{t,1})$ represents the covariance between an investor's inter-temporal rate of substitution, $\widetilde{m}_{t,1}$, and the random future price of the investment at $t + 1$, $\widetilde{P}_{t+1,s-1}$, based on the information available to investors today (t). The subscript is reduced by one because an investment with time to maturity s at time t becomes an investment with time to maturity $s - 1$ at time $t + 1$ (Cochrane 2005).

Equation 6 expresses the value of a risky asset as the sum of two terms. The first term is the asset's expected future price discounted at the risk-free rate. It may be called the risk-neutral present value because it represents a risky asset's value if investors did not require compensation for bearing risk (notice the parallel with the fundamental pricing equation, Equation 1, if it had one cash flow and no risk premium). In Example 5, this value is 0.867234.

The covariance term is the discount for risk. Note that with a one-period default-free bond, the covariance term is zero because the future price is a known constant ($1) and the covariance of a random quantity with a constant is zero; and intuitively, its value is given by the first term. Consequently, Equation 6 reduces to Equation 3 for the one-period default-free bond. But with the two-period default-free bond, the future price of $1 two periods in the future is known with certainty, but the price one period in the future is not. Consequently, the covariance term is not zero.

In general with risk-averse investors, the covariance term for most risky assets is expected to be negative. That is, when the expected future price of the investment is high, the marginal utility of future consumption relative to that of current consumption is low. Alternatively, during bad economic times, investors expect a smaller labor income in the future, so the marginal utility of future consumption, and hence the inter-temporal rate of substitution, is higher. This relationship leads investors to demand a higher required rate of trade-off of future for current consumption—as in bad economic times when the labor market contracts. Bad economic times also tend to be associated with declining risky asset payouts (declining earnings and dividends for ordinary shares and defaults for bonds), leading to declining asset prices. The result is that the covariance term for risky assets is typically negative, so the price of the asset is lower. This negative covariance term results in a positive risk premium, $\rho^i_{t,s}$, in Equation 1 because a lower price today leads to a higher return over time. Holding all else constant, the risk premium term and the required return for an asset should be higher, and its current market price is lower the larger the magnitude of the negative covariance term.

EXAMPLE 6

An Alternative Method to Evaluate the Price Discount for Risk

1. The covariance between the investor's willingness to invest and the price of the two-year bond next period can also be computed as follows (recall the standard formula for covariance is $\text{cov}(\bar{x}, \bar{y}) = \Sigma p_i[x_i - E(x)][y_i - E(y)]$:

   $\text{cov}_t(\widetilde{m}_{t,1}, \widetilde{P}_{t+1,2-1}) = 0.4(0.954676 - 0.954498) \times (0.839181 - 0.908576) + 0.6(0.954379 - 0.954498) \times (0.954840 - 0.908576)$

$= -0.000008$

In the bad state of the economy, the willingness to invest (0.954676) is above its average (0.954498), yet the bond price (0.839181) is below its average (0.908576). The reverse is true in the good state. Thus, the covariance between the inter-temporal rate of substitution and the price of the asset is negative. This result means the investor finds this investment inferior to one with a payoff that is independent of her willingness to invest. In particular, we have

$$P_{t,s} = \frac{E_t\left(\tilde{P}_{t+1,s-1}\right)}{1 + l_{t,1}} = -0.000008.$$

With a lower price, the return on the two-year bond is higher.

In this example, the price discount is not too large because the risk between a one- and two-year US government bond is not that crucial. However, riskier assets, such as equity, will have a higher discount. In addition, the higher risk premium on equity still follows from the covariance between the cash flow and the investor's willingness to invest over the time horizon of the investment. Thus, a higher risk premium for stocks arises from a larger value for this covariance.

The risk premium can be computed as follows: The expected holding period return on the s period bond through time $t + 1$, using the results of Example 5, is given by

$$\begin{aligned} r_{t,s} &= \frac{E_t\left(\tilde{P}_{t+1,s-1}\right) - P_{t,s}}{P_{t,s}} \\ &= \frac{0.908576 - 0.867226}{0.867226} = 0.047681, \text{ or } 4.7681\%, \end{aligned} \tag{7}$$

so the risk premium $\rho^i_{t,s} = r_{t,s} - l_{t,1} = 0.047681 - 0.047671 = 0.00001$. Alternatively, Equation 7 and 6 can be manipulated so that

$$\begin{aligned} r_{t,s} - l_{t,1} &= \frac{E_t\left(\tilde{P}_{t+1,s-1}\right)}{P_{t,s}} - \left(1 + l_{t,1}\right) = -\frac{\left(1 + l_{t,1}\right)}{P_{t,s}}\mathrm{cov}_t\left(\tilde{m}_{t,1}, \tilde{P}_{t+1,s-1}\right) \\ &= -\left(1 + l_{t,s}\right)\mathrm{cov}_t\left(\tilde{m}_{t,1}, \frac{\tilde{P}_{t+1,s-1}}{P_{t,s}}\right) = -\frac{(1 + 0.047671)}{0.867226} \times (-0.000008) \\ &= 0.00001 = \rho^i_{t,s}, \end{aligned} \tag{8}$$

which is the return premium demanded by investors because of the uncertain Time 1 price of the riskless two-period bond.[5]

5 Notice that simultaneously multiplying Equation 6 by $(1 + l_{t,1})$ and dividing by $P_{t,s}$ gives

$$(1 + l_{t,1}) = \frac{E_t\left(\tilde{P}_{t+1,s-1}\right)}{P_{t,s}} + \left(1 + l_{t,1}\right)\mathrm{cov}_t\left(\tilde{m}_{t,1}, \frac{\tilde{P}_{t+1,s-1}}{P_{t,s}}\right),$$

so we can write

$$\frac{E_t\left(\tilde{P}_{t+1,s-1}\right)}{P_{t,s}} - \left(1 + l_{t,1}\right)$$

in Equation 8 as

$$\frac{E_t\left(\tilde{P}_{t+1,s-1}\right)}{P_{t,s}\left(\tilde{m}_{t,1}, \frac{\tilde{P}_{t+1,s-1}}{P_{t,s}}\right)} - \frac{E_t\left(\tilde{P}_{t+1,s-1}\right)}{P_{t,s}} - \left(1 + l_{t,1}\right)\mathrm{cov}_t\left(\tilde{m}_{t,1}, \frac{\tilde{P}_{t+1,s-1}}{P_{t,s}}\right) = -\left(1 + l_{t,1}\right)\mathrm{cov}_t$$

This relationship implies that an asset's risk premium, $\rho^{i}_{t,s}$ in Equation 1, is driven by the covariance of its returns with the inter-temporal rate of substitution for consumption and can exist even for a default-free bond because of the uncertainty of its price before maturity. Most risky assets have returns that tend to be high during good times, when the marginal value of consumption is low, and low during bad times, when the marginal value of consumption is high, and so bear a positive risk premium. Any asset that tended to have relatively high returns when the marginal utility of consumption was high would provide a type of hedge against bad times, bear a negative risk premium, and have a relatively high price and low required rate of return.

7 DEFAULT-FREE INTEREST RATES AND ECONOMIC GROWTH

☐ explain the relationship between the long-term growth rate of the economy, the volatility of the growth rate, and the average level of real short-term interest rates

From the previous discussion, it is a relatively small conceptual step to understand the relationship between an economy's GDP growth and real default-free interest rates. If there is a known independent change in real GDP growth or a change that can be forecasted perfectly, then an increase in real GDP growth should lead to an increase in the real default-free rate of interest because more goods and services will be available in the future relative to today. The result is that investors' willingness to substitute over time will fall, resulting in less saving and more borrowing, so that the real default-free interest rate increases, as in Equation 4.

But GDP growth from one period to the next cannot be perfectly anticipated. Under these uncertain circumstances, interest rates will still be positively related to the expected growth rate of GDP, but additionally they will be positively related to the expected volatility of GDP growth.

EXAMPLE 7

The Effect of Volatility on Prices

1. One can see the effect of volatility by doubling the standard deviation of the random variable $\tilde{z}$ from what was assumed in Example 1. In this case, the price of the one-period bond in Example 1 would be

$$E_t\left(\widetilde{m}_{t,1}\right) = 0.4 \times 0.954855 + 0.6 \times 0.954260 = 0.954498.$$

Notice that the expected value is the same as in Example 1 but that the prices in each state are more dispersed, reflecting the doubling of the standard deviation. For the two-period default-free bond, continuing with the parameter values (a' and b') from Example 5, we would compute

$$E_t\left(\widetilde{P}_{t+1,2-1}\right) = 0.4 \times 0.776625 + 0.6 \times 1.005451 = 0.913921.$$

Notice that doubling the volatility leads to a somewhat unrealistic price greater than 1 (implying a negative yield) in the good state, even though the expected price is less than 1.

Then,

$$P_{t,s} = E_t\left(\widetilde{P}_{t+1,2-1}\widetilde{m}_{t,1}\right)$$
$$= 0.4 \times 0.776625 \times 0.954855 + 0.6 \times 1.005451 \times 0.954260 = 0.872303.$$

The risk neutral price is $\frac{E_t\left(\widetilde{P}_{t+1,2-1}\right)}{1 + l_{t,1}} = \frac{0.913921}{1.047671} = 0.872336$. So, from Equation 6, $\text{cov}_t\left(\widetilde{m}_{t,1}, \widetilde{P}_{t+1,2-1}\right) = 0.872303 - 0.872336 = -0.000033$.
As a result, the holding period return on a two-period bond for one year is higher [i.e., (0.913921 − 0.872303)/0.872303 = 4.771%, compared with 4.768% in Example 6], and because of the higher volatility, investors require a higher premium.

There are two practical implications of this analysis for the values of real default-free interest rates:

- An economy with higher trend real economic growth, other things being equal, should have higher real default-free interest rates than an economy with lower trend growth. We should thus expect to find that real default-free interest rates in fast-growing, developing economies, such as India and China, are higher than in slower-growing, developed economies, such as Western Europe, Japan, and the United States. The higher rate of economic growth occurs for developing economies because a developing economy is typically below its steady state growth, so it grows faster to catch up. During these periods, the marginal product of capital (the additional output resulting from the addition of one unit of capital, holding all else constant) would be expected to be higher, so the real default-free interest rate should also be expected to be higher. Of course, this advantage will dissipate as the economy matures, as in the case of Japan and Western Europe from 1950 to 2000.
- Again, other things being equal, real interest rates are higher in an economy in which GDP growth is more volatile than in an economy in which growth is more stable.

EXAMPLE 8

The Present Value Model and Macroeconomic Factors

1. An asset's risk premium is high when:

   A. there is no relationship between its future payoff and investors' marginal utility from future consumption.

B. there is a positive relationship between its future payoff and investors' marginal utility from future consumption.

C. there is a negative relationship between its future payoff and investors' marginal utility from future consumption.

Solution:

C is correct. An asset's risk premium is determined by the relationship between its future payoff and the marginal value of consumption as given by the covariance between the two quantities. When the covariance is negative—that is, payoffs are low and expected utility from consumption is high—or equivalently, when times in the future are expected to be bad and the value of an extra unit of consumption is high, the risk premium will be high. When the covariance term is zero (there is no relationship), the asset is risk free. When the covariance term is positive, the asset is a hedge and will have a rate of return less than the risk-free rate.

2. The relationship between the real risk-free interest rate and real GDP growth is:

   A. negative.

   B. neutral.

   C. positive.

Solution:

C is correct. The real risk-free rate is positively related to real GDP growth. An increase in real GDP growth reduces the need for investors to save for future consumption because more goods and services will be available to them in the future relative to today as a result of higher expected income in the future. A higher real rate of interest is needed to induce individuals to save for future consumption in such circumstances.

3. The relationship between the real risk-free interest rate and the volatility of real GDP growth is:

   A. negative.

   B. neutral.

   C. positive.

Solution:

C is correct. The real risk-free rate is positively related to the volatility of real GDP growth. An increase in volatility of real GDP growth means that there is greater risk that the income available for consumption will be lower than expected. Therefore, risk-averse investors will require a higher real rate of return in compensation.

4. A risky asset offers high positive returns during business downturns. A colleague argues that the nominal required rate of return on the asset may be less than the nominal risk-free rate. Is the colleague correct?

   A. Yes

   B. No, the return must be higher than the nominal risk-free rate.

   C. No, the relationship between the asset's nominal return and the nominal risk-free rate is indeterminate.

Solution:

A is correct. For the required return to be less than the risk-free rate, the asset's risk premium would need to be negative. Because the asset supplies relatively high returns in economic conditions in which the marginal utility of consumption is relatively high, the covariance term in Equation 6 is positive and the asset thus bears a negative risk premium.

5. Suppose you are analyzing the expected impact of an increase in real GDP growth above trend on overall equity market valuation. Assume real growth in income of the corporate sector follows real GDP growth. Assume also that there is no impact of the increase on inflation. On the basis of theory and holding all else constant, explain why the impact of the assumed increase in real GDP growth on overall equity market valuation is ambiguous.

Solution:

Equation 1 can be applied to the overall equity market, which is the aggregate of individual equity securities. The increase in real GDP growth would be expected to affect both the numerator and denominator of Equation 1 in offsetting ways, so the overall effect on equity market value is ambiguous. The impact of an increase in real GDP growth on expected corporate earnings is positive by assumption, which by itself would suggest an increase in equity market value by a larger numerator value in Equation 1. However, the increase in real GDP should also increase the real risk-free rate, which by itself would suggest a decrease in equity market value by increasing the rate at which expected cash flows are discounted. We cannot infer which effect will dominate from the information given. The overall effect on equity market value is ambiguous, under the assumptions given.

8 REAL DEFAULT-FREE INTEREST RATES AND THE BUSINESS CYCLE

☐ explain how the phase of the business cycle affects policy and short-term interest rates, the slope of the term structure of interest rates, and the relative performance of bonds of differing maturities

One of the crucial insights that macro-finance provides is that there should be a connection between the real risk-free rate of interest available in an economy and the underlying trend rate of economic growth in the same economy. We explored the roots of this relationship earlier. To recap briefly, the willingness of investors to substitute future wealth for current consumption will be inversely related to the change in real GDP growth. In a world where GDP growth could be forecasted perfectly, there will be a positive relationship between the real risk-free rate and real GDP growth. But GDP growth is not perfectly predictable. Because of this unpredictability, we also concluded that the real risk-free rate would not only be positively related to real GDP growth but also positively related to the volatility of real GDP growth.

Equation 1 shows that the real default-free required return, $l_{t,s}$, is a component of the discount rate that we apply to the cash flows generated by all financial instruments.

Economic Growth and Real Yields

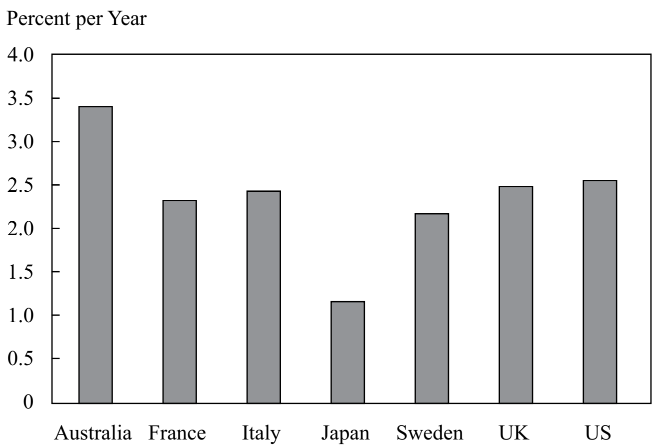
For evidence of the relationship between real interest rates and GDP growth, we could focus on the yields available from inflation-linked bonds issued by governments in developed economies. These bonds pay a "real" return (or yield) plus a return that is linked directly to an index of consumer prices. Index-linked bonds are issued by many governments in developed economies, including Canada, France, Germany, Italy, Sweden, the United Kingdom, and the United States, and also by some governments in developing countries, such as Brazil. In some markets—for example, the United Kingdom's index-linked gilt market—both the coupon and principal payments from these bonds are indexed to a measure of consumer prices. In other markets—for example, the US Treasury Inflation-Protected Securities (TIPS) market—the principal payment is indexed and the coupon is a function of the indexed principal. In both cases, any increase in the level of the consumer price index over time (that is, positive inflation) leads to an increase in both the coupon payment and eventual principal payment. Although the details of the indexation vary from bond market to bond market, for all practical purposes, we can think of these bonds as being inflation protected.

Given the earlier discussion in sections on real default-free interest rates and default-free interest rates and economic growth, other things being equal, we would expect the (real) yields on inflation-indexed bonds to be higher for those countries with high growth, such as India and China, relative to those issued by, say, the UK or US governments, where economic growth is much lower. Other things being equal, we should also expect to see real yields on short-dated index-linked bonds issued by governments of economies that are very volatile to be higher than real yields on those issued by governments of economies that are less volatile.

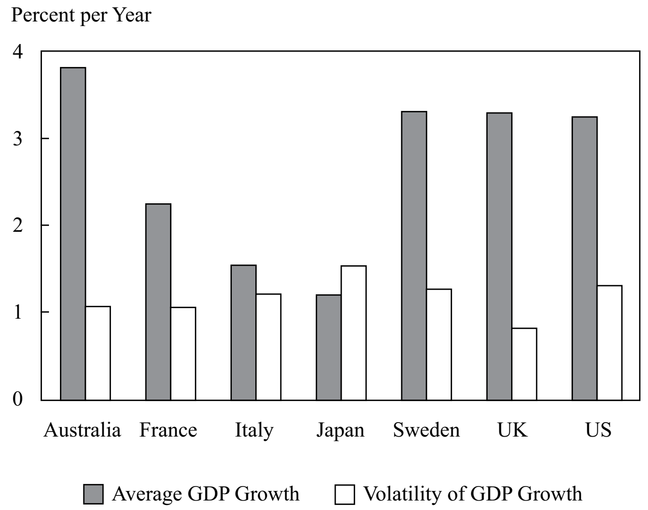
Although many index-linked government bond markets are relatively new, we can examine the cross-sectional relationship between economic growth and real risk-free yields relatively easily. Panel A of Exhibit 1 shows the real yields on a set of short-dated index-linked government bonds in 2007, immediately prior to the 2008–2009 global financial crisis. The real yield on short-dated Japanese government bonds at that time were lower than elsewhere, whereas Panel B shows that Japanese growth had been historically very low and not very volatile up to that point in time. Of the developed-economy bond yields in this exhibit, those issued by the Australian government offered the highest yield, perhaps reflecting the relatively strong Australian economic growth shown in Panel B.

Exhibit 1: Real Yields, GDP Growth, and Volatility for Various Countries

Real Yields, July 2007



Growth and Volatility, 1996–2007



Sources: Based on data from Thomson Reuters and the authors' calculations.

It is difficult to discern a very clear pattern between historical economic growth, the volatility in that growth, and short-term real yields. Nonetheless, it is interesting to note that the correlation between this limited set of bond yields (Panel A of Exhibit 1) and historical growth (gray bars in Panel B of Exhibit 1) is 0.57, but the correlation with historical volatility (white bars in Panel B) is 0.74. So there does appear to be some support for the prediction of macro-theory, although this sample is of course very limited. One of the reasons why there is perhaps not a clearer relationship is that the real yield data are forward looking. The real yield data represent the required real return on these bonds based on expected future growth and volatility in that growth, whereas the GDP-based variables represent historical growth and volatility. If investors use the past as a guide for the future, they might expect a reasonably high correlation between past growth and current real yields. But the past may be a very bad guide for the future, particularly in the case of rapidly developing countries or following the sort of major shock to global economic growth that occurred following the collapse of Lehman Brothers.

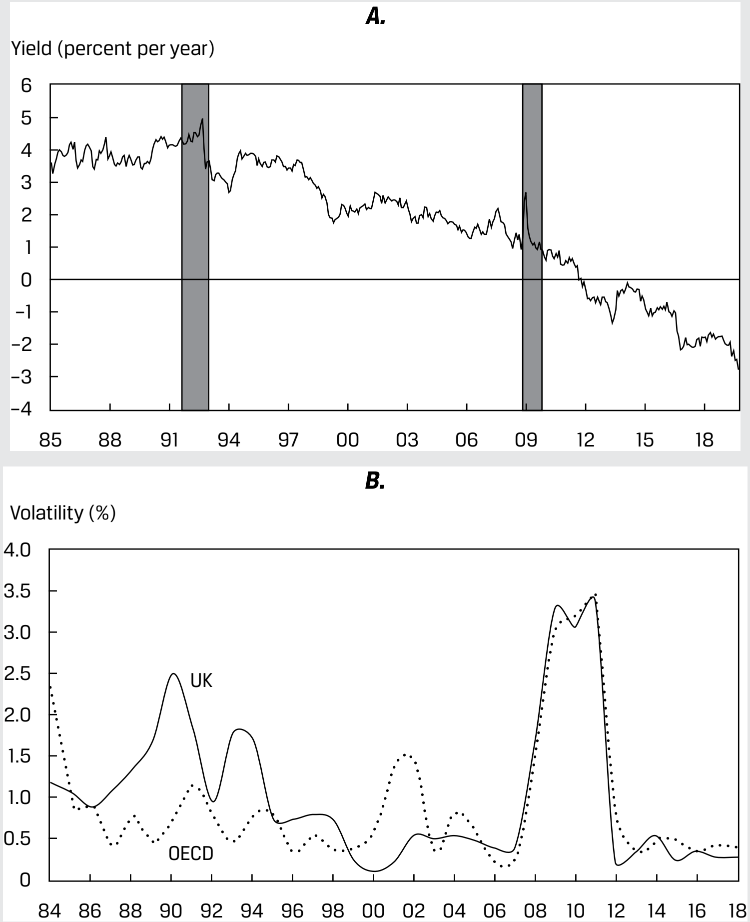
EXAMPLE 9

The Evolution of Real Yields

1. We have seen that there is at least tentative evidence to suggest a positive relationship between economic growth and the yields on short-dated real government bonds. But how does the business cycle affect these yields over time?

   The problem in gathering evidence on the drivers of real yields over time is that index-linked government bond markets are a fairly recent financial innovation, especially when compared with conventional government bonds. For example, it was only in 1997 that the first index-linked bond was issued by the US Treasury. The oldest index-linked government bond market is the United Kingdom's. The UK government issued its first index-linked gilt in 1981. To investigate the connections between the macroeconomy and the real risk-free rate over time, we can focus on the UK's index-linked gilt market.

Exhibit 2: Real Yields on UK Index-Linked Gilt and Volatility of UK and OECD GDP Growth



Notes: GDP growth is represented by a three-year moving standard deviation of the variable. Shaded areas in Panel A indicate UK recessions.

Sources: Based on data from OECD.Stat and UK National Statistics Office.

Panel A of Exhibit 2 shows the real yield on a short-dated constant maturity UK index-linked gilt. We will begin by focusing on the period from 1985 to 2007. Although fairly volatile, there is a clear downward trend in this yield from 1985 to 2007. One explanation for this decline could be a commensurate decline in expectations about UK economic growth. However, real economic growth between 1985 and 1999 averaged 2.8%, and between 2000 and 2007 it averaged 2.7%. There was very little change in average growth at a time when real yields were falling, and therefore, it is probably fair to assume that expectations of future growth were relatively stable over this time too. Panel B of Exhibit 2 shows the volatility of UK real GDP growth as represented by a three-year moving standard deviation of this variable. It shows that the volatility of UK economic growth declined quite dramatically from 1995 to 2007. This decline in UK GDP volatility was also experienced elsewhere in the global economy. The same chart shows the decline in GDP volatility for OECD countries. This decline in economic volatility has been called "the great moderation"—that is, a period when the global economy and its financial markets were characterized by relatively low levels of volatility. Therefore, one plausible explanation for the declining level of real interest rates in the United Kingdom is that they were driven down by the moderation in economic volatility between the early 1990s and 2007.

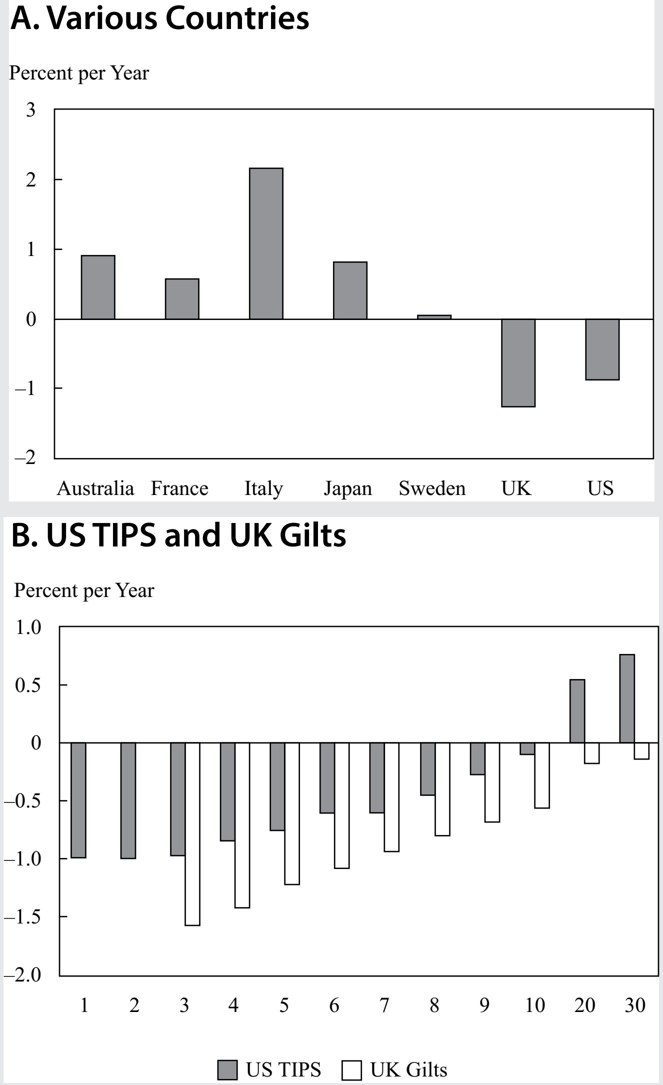
The evidence in Exhibit 2 suggests that declining levels of economic volatility led to declining levels of the real default-free interest rate in the United Kingdom between 1999 and 2007. However, the absence of such markets elsewhere over this sample period does not mean that the same phenomenon was absent or irrelevant in other developed-economy bond markets. The yield on a conventional government bond includes a number of components, one of which is the real default-free rate of return. So, in all likelihood, declining global economic volatility led to declines in the real rates of return required by investors elsewhere, which, in turn, may have contributed to the decline in conventional government bond yields. See Example 10 for an illustration of how the global financial crisis affected real short rates. We will focus on the drivers of conventional government bond yields relative to index-linked government bond yields at a later stage.

EXAMPLE 10

Post–Global Financial Crisis, 2008–2011 Real Default-Free Yields

1. Exhibit 3 shows the yields on short-dated index-linked bonds at the end of 2011. Compared with their pre-crisis levels shown in Exhibit 2, they had all fallen. The collapse of Lehman Brothers and the ensuing liquidity and sovereign debt crisis caused economic and financial market volatility to rise substantially, as shown in Panel B of Exhibit 2. Other things being equal, one would have expected the real yields to rise, not to fall. But other things were not equal. One explanation for the fall in these real yields is that, despite the higher volatility experienced in 2008–2011, investors believed that future real economic growth would be lower and, therefore, that the equilibrium real yield in these economies was deemed to be commensurately lower.

Exhibit 3: Real Yields on Short-Dated Index-Linked Bonds, December 2011



Sources: Based on data from Bloomberg and Thomson Reuters.

Panel A of Exhibit 3 also shows that real short-dated rates in the United Kingdom and United States were not just lower than they were in the immediate pre-crisis period but that they were negative, too. Panel B shows that most inflation-linked bonds issued by the US government (TIPS) and all inflation-linked bonds issued by the UK government (index-linked gilts) at this time were also negative (term structure will be explored in more detail later). One explanation for the lower yields may have been the fear among some investors of very high levels of inflation in the future. Arguably, the easy monetary policy, including both formal (US and UK) and informal (eurozone) quantitative easing programs, implemented because of the collapse of Lehman Brothers along with other events, may have led enough investors to believe that inflation-linked government bonds offered the inflation protection they needed despite the low and even negative historical real yields.

But it is also important to remember that index-linked bonds issued by developed-economy governments not only are quite special, given the credit and inflation protection that they provide, but also are often in very limited supply. These characteristics mean that in times of crisis and of great uncertainty, investors may see them as a safe haven for their capital, which can, in turn, drive down their yields.

Real Default-Free Interest Rate Summary

The real default-free interest rate, which we have proxied here with the yields on short-term inflation-protected government bonds, has a close connection with the business cycle via the related connection with the saving decisions of individuals. We can put this discussion in the context of the basic pricing shown in Equation 1. For a real default-free bond, Equation 1 simplifies to

$$P_t^i = \sum_{s=1}^{N} \frac{CF_{t+s}^i}{\left(1 + l_{t,s}\right)^s}. \quad (9)$$

Because it is a fixed-interest investment that is default-free, the cash flow at time $t + s$ is certain. Equation 9 implies that it is only changes in $l_{t,s}$ that will affect the price of such a bond. In turn, $l_{t,s}$ will be determined by real economic growth and the volatility in economic growth over time as a result of the aggregation of the consumption and saving decisions of individual investors.

9 THE YIELD CURVE AND THE BUSINESS CYCLE

☐ explain how the phase of the business cycle affects policy and short-term interest rates, the slope of the term structure of interest rates, and the relative performance of bonds of differing maturities

Earlier, we considered the determination of the price of a real default-free bond (see Equation 9). The analysis demonstrated that the saving and investment decisions of investors mean that the expected return on these bonds will be positively related to both expected real GDP growth and the expected volatility of this growth. We now move on to consider the price of a default-free bond that pays a fixed nominal (currency) amount when it matures. We will consider, for example, a bond issued by a government in a developed economy where the prospect of default is so negligible that it is ignored.

What factors would affect the price of such a bond? First, we consider a world without inflation. In this world, investors would still be giving up current consumption by investing in this bond today, in which case Equation 9 would be appropriate. But of course, deferring consumption at time t in a world with positive inflation will have an impact on the quantity of goods that can be bought at time $t + s$ when the bond matures. Investors will want to be compensated by this bond for the inflation that they expect between t and $t + s$, which we define as $\theta_{t,s}$. If investors could forecast inflation perfectly, they would demand a return given by $l_{t,s} + \theta_{t,s}$ to compensate them for the expected inflation and ensure the real level of consumption. But unless the investment horizon is very short, investors are unlikely to be very confident in their ability to forecast inflation accurately. Because we generally assume that investors are risk averse and thus need to be compensated for taking on risk and because they seek compensation for expected inflation, they will also seek compensation for taking on the uncertainty related to future inflation. We denote this risk premium by $\pi_{t,s}$, which is distinct from the risk premium in Equation 1 ($\rho^i_{t,s}$).[6] We can rewrite our basic pricing formula for a default-free nominal coupon-paying bond from Equation 9 as follows:

$$P^i_t = \sum_{s=1}^{N} \frac{CF^i_{t+s}}{\left(1 + l_{t,s} + \theta_{t,s} + \pi_{t,s}\right)^s}. \quad (10)$$

Note that the bond's payoff is still certain in nominal terms because we are assuming that there is a negligible chance that the issuer (a developed-economy government) will default on its commitments. It is the real value of this payoff that is now uncertain, hence the need for a risk premium, $\pi_{t,s}$, on nominal bonds.

EXAMPLE 11

The Risk Premium for Inflation Uncertainty

1. Suppose that an analyst estimates that the real risk-free rate is 1.25% and that average inflation over the next year will be 2.5%. If the analyst observes the price of a default-free bond with a face value of £100 and one full year to maturity as being equal to £95.92, what would be the implied premium embedded in the bond's price for inflation uncertainty?

Solution

The (approximate) implied premium can be calculated as follows:

$$\pi_{t,s} = 0.504\% = \frac{100}{95.92} - (1 + 0.0125 + 0.025).$$

Having established Equation 10, we will now focus on the relationship between short-term nominal interest rates and the business cycle, in which the nominal bond issued by a government in a developed economy has a very short maturity—for example, a US government Treasury bill.

Short-Term Nominal Interest Rates and the Business Cycle

Treasury bills (T-bills) are very short-dated nominal zero-coupon government bonds. T-bills are issued by most developed-economy governments or by their agents to help smooth the cash flow needs of the government. The short-dated nature of T-bills and the fact that they are often used to implement monetary policy means that their yields are also usually very closely related to the central bank's policy rate. Indeed, because of their short-dated nature, the uncertainty that investors would have about inflation over an investment horizon of, say, s equals three months will usually be relatively low. Therefore, for the purposes of the exposition in this section of the reading, we will assume that $\pi_{t,s}$ is so negligible that we can ignore it. So, we can modify Equation 10 to give Equation 11, which can capture all of the salient features of the pricing dynamics of a T-bill:

$$P^i_t = \frac{CF^i_{t+s}}{\left(1 + l_{t,s} + \theta_{t,s}\right)^s}. \quad (11)$$

Note that the summation term is not needed because there is only one payment from a T-bill.

6 Even though it is a risk premium, we have suppressed the superscript i on the inflation uncertainty risk premium because it is not asset specific and applies across all asset classes.

We have already examined the way in which the real default-free rate of interest, $l_{t,s}$, will vary over time with the business cycle and how it may also be affected by its status as a haven in times of economic uncertainty. We now move on to consider how a central bank's policy rate, which is a short-term nominal interest rate, evolves with the business cycle.

10 TREASURY BILLS AND THE BUSINESS CYCLE

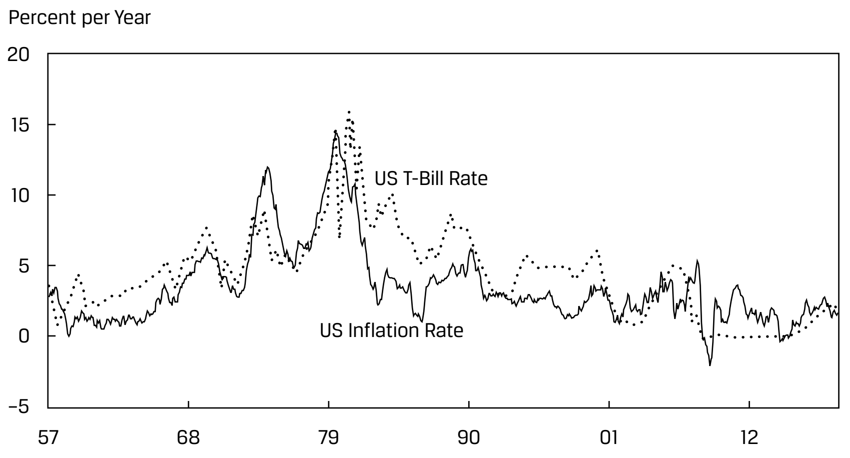
☐ explain how the phase of the business cycle affects policy and short-term interest rates, the slope of the term structure of interest rates, and the relative performance of bonds of differing maturities

To summarize briefly, the nominal rate of interest will equal the real interest rate that is required to balance the requirements of savers and investors plus investors' expectations of inflation over the relevant borrowing or lending period. It follows that short-term nominal interest rates will be positively related to short-term real interest rates and to short-term inflation expectations. Other things being equal, we would also expect these interest rates to be higher in economies with higher, more volatile growth and with higher average levels of inflation over time.

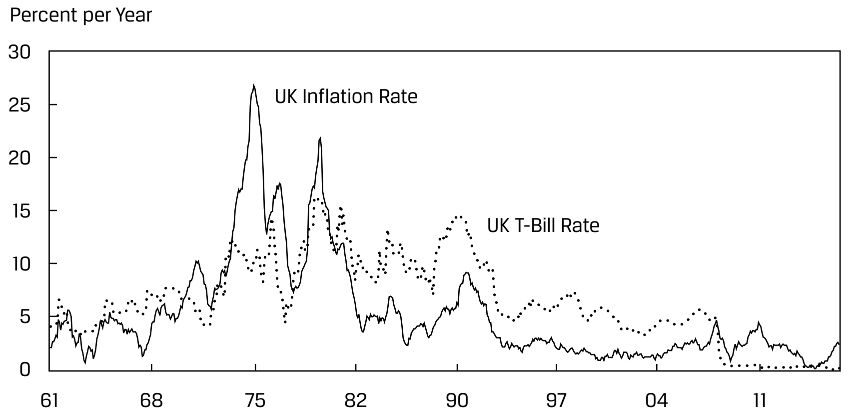
Panel A of Exhibit 4 shows the yield on a three-month US T-bill, and Panel B shows the yield on an equivalent T-bill issued by the UK government. In each panel of the exhibit, we also present the inflation rates in these two economies. There is a close correlation between measured inflation and T-bill yields in both economies. Although measured inflation is not the same as expected inflation, it is likely that current inflation plays a big role in the formation of inflation expectations, particularly over the very short investment horizon involved when investing in a T-bill.

Exhibit 4: Treasury Bill Rates and Inflation

United States



United Kingdom

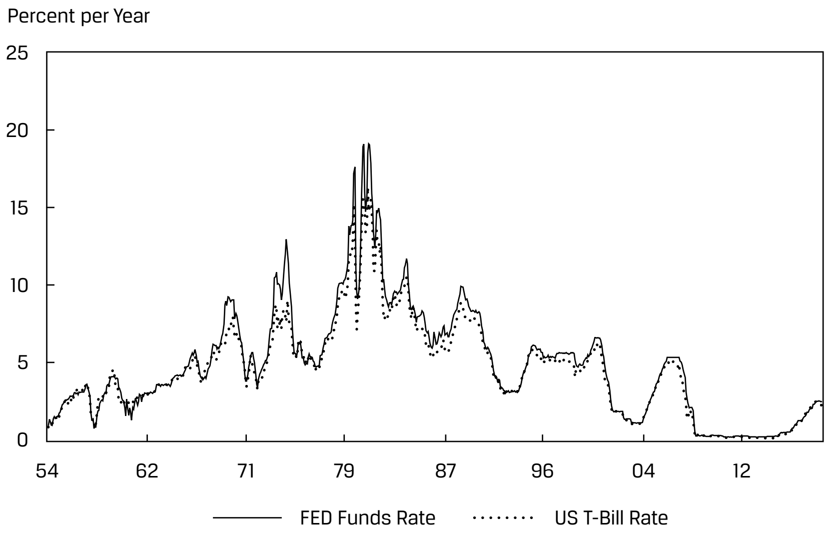


Source: Based on data from FRED.

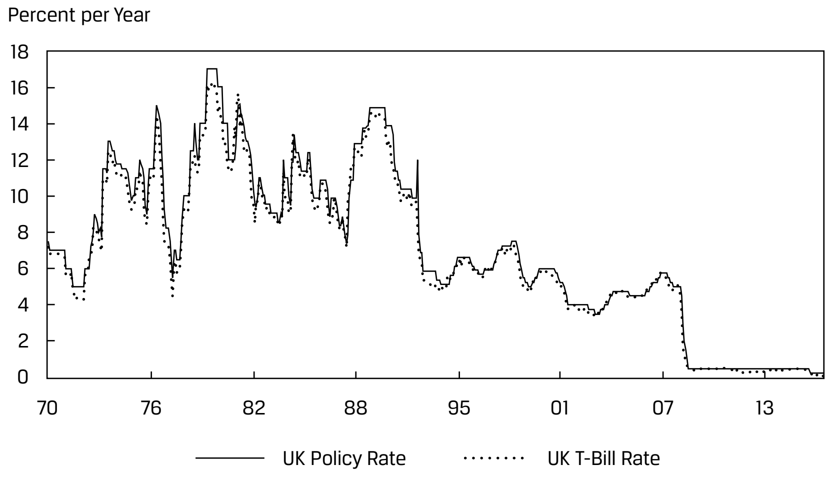
Exhibit 4 clearly shows that the inflation environment is a key driver of short-term interest rates. The central banks and monetary authorities responsible for setting interest rates in an economy do so in response to the economy's position in the business cycle—cutting their policy rates when activity and/or inflation are judged to be "too low" and raising rates when activity and/or inflation are judged to be "too high." In other words, a responsible central bank or monetary authority will usually set its policy rate with reference to the level of expected economic activity and the expected rate of increase of prices—that is, inflation. Exhibit 5 shows the close relationship between the yields on short-term default-free T-Bills in the United States and United Kingdom and the policy rates of their respective central banks.

Exhibit 5: Interest Rates and Policy Rates

United States



United Kingdom



Source: Based on data from Thomson Reuters.

A US economist, John Taylor, devised a rule for setting policy rates, a rule that could help rate setters gauge whether their policy rate is at an "appropriate" level (Taylor 1993). This rule is known as the Taylor rule, and it takes the following form:

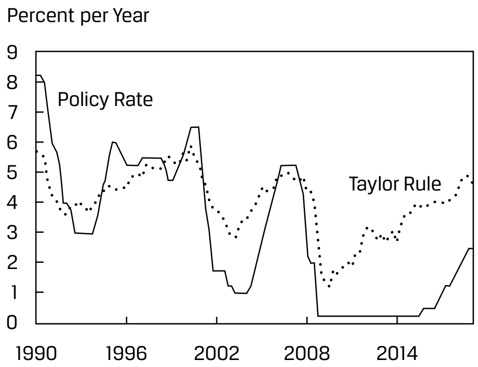
$$\begin{aligned} pr_t &= l_t + \iota_t + 0.5\left(\iota_t - \iota_t^*\right) + 0.5\left(Y_t - Y_t^*\right) \\ &= l_t + 1.5\,\iota_t - 0.5\,\iota_t^* + 0.5\left(Y_t - Y_t^*\right), \end{aligned} \quad (12)$$

where pr_t is the policy rate at time t, l_t is the level of real short-term interest rates that balances long-term savings and borrowing in the economy, ι_t is the rate of inflation, ι_t^* is the target rate of inflation, and Y_t and Y_t^* are, respectively, the logarithmic levels of actual and potential real GDP. The difference between Y_t and Y_t^* is known as the "output gap," which is essentially measured in percentage terms. When the output gap is positive, it implies that the economy is producing beyond its sustainable capacity. This situation is similar to a marathon runner who sets off way too fast at the start of a race; in the end, he will overheat and break down unless he reduces his running pace. Conversely, when the output gap is negative, it implies that the economy is producing below its sustainable capacity. This situation is similar to a marathon runner who sets off too slowly. If he wants to win the race, at some point, he will have to use up conserved energy and speed up. Positive output gaps are usually associated with high and/or rising inflation, whereas negative output gaps are usually accompanied by high levels of unemployment. Generally, the policy rule should have a larger weight on inflation (1.5) relative to the weight on output (0.5). The purpose is to stabilize inflation over the longer term near the targeted inflation rate (note that the reason for the weightings is that the inflation rate appears twice in the equation; see the first line of Equation 12). When inflation is close to the targeted or preferred rate and when the output gap is zero, the appropriate policy rate will be equal to the level of the short-term real interest rate, l_t, that balances long-term savings and borrowing in the economy plus the targeted/preferred rate of inflation. This level of the policy rate is often referred to by economists as the neutral policy rate—that is, the policy rate that neither spurs on nor impedes real economic activity. Other things being equal, when inflation is above (below) the targeted level, the policy rate should be above (below) the neutral rate, and when the output gap is positive (negative), the policy rate should also be above (below) the neutral rate. For example, if l_t is 2.0%, ι_t is 3.0%, ι_t^* is 2.0%, and the output gap is 2.0%, then the "appropriate" policy rate implied in the Taylor rule would be 6.5%.

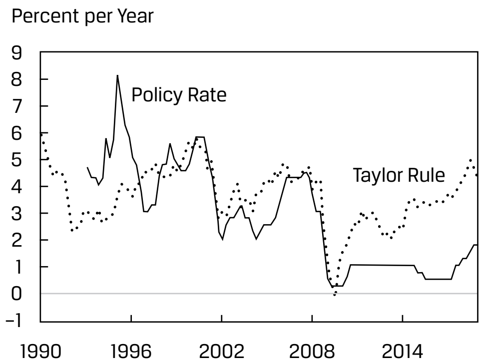
Using fairly conservative parameters, including inflation targets when they are known, and a measure of the output gap estimated by the OECD, we have calculated policy rates based on the Taylor rule for three developed economies back to 1990, as shown in Exhibit 6. The policy rates based on the Taylor rule for the United States, shown in Panel A, seem to track the Fed's actual policy rates fairly closely until the collapse of the high-tech bubble in the early 2000s. According to the Taylor rule, the Fed kept policy rates "too low for too long" between 2002 and 2005. A similar picture emerges for Canada, as shown in Panel B. There is less evidence that policy rates were kept "too low for too long" after the collapse of the high-tech bubble in the United Kingdom, as shown in Panel C. More recently, in response to the liquidity and credit crisis, all three central banks cut their policy rates sharply. According to the Taylor rule, for all three economies, policy rates were "too low" by the end of 2018.

Exhibit 6: Policy Rates and Taylor Rule Calculations

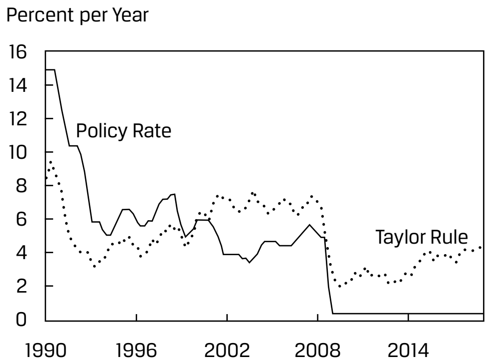
United States



Canada



United Kingdom



Sources: Based on data from each country's central bank and the authors' calculations.

The notion that short-term interest rates can be at too high or too low a level implies that the relationship between short-term interest rates and the business cycle are interdependent. In other words, instead of moderating the business cycle over time with deft changes to the policy rate, central bankers can exaggerate any cycle by not responding optimally to economic conditions—that is, by committing policy errors. For example, setting rates too low for too long risks creating a credit bubble, whereas setting them too high for too long could lead to recessionary or even depression-like economic conditions.

Short-Term Interest Rate Summary

Short-term default-free interest rates tend to be very heavily influenced by the inflation environment and inflation expectations over time. But they will also be influenced by real economic activity, which, in turn, is influenced by the saving and investment decisions of households. But these interest rates will also be affected by the central bank's policy rate, which, in turn, should fluctuate around the neutral policy rate as central banks respond over time to deviations in inflation from a preferred or target rate and to developments in the output gap. Finally, it is important to remember that the neutral rate will also vary with the level of real economic growth and with the expected volatility of that growth. In addition, the neutral rate might also change if the level of inflation targeted or preferred by the central bank changes. Between 1992 and 1997, the UK inflation target was between 2% and 4%; in 1997, the target became 2.5% with a 1 percentage point allowance around this target. The target was changed in 2003 to 2.0% and to a different definition of inflation, with a 1 percentage point allowance around this target.

In the next section of this reading, we will focus on the relationship between the underlying economy and longer-term nominal default-free government bonds.

11 CONVENTIONAL GOVERNMENT BONDS AND BREAK-EVEN INFLATION RATES

☐ describe the factors that affect yield spreads between non-inflation-adjusted and inflation-indexed bonds

The pricing equation shown in Equation 10 can be used to highlight the key components that go into pricing conventional (coupon-paying) government bonds. We will consider the impact that time to maturity can have on the pricing formula later, but first we will focus on the impact of inflation expectations on conventional bond prices.

Break-Even Inflation Rates

The fundamental difference between the pricing formula as applied to, for example, a three-month T-bill (as in Equation 11) and its application to, for example, a default-free zero-coupon bond (as in Equation 10) relates to their investment horizons. The relative certainty about the real payoff from a three-month T-Bill and thus the relative certainty about the amount of consumption that the investor will be able to undertake with the payoff means that the investment in the T-Bill will be a good hedge against possible bad consumption outcomes. In other words, the payoff, in real terms, from a three-month T-bill is highly unlikely to fall if the investor loses his or her job during the T-bill's three-month investment horizon. The low, probably zero, correlation between the T-bill's payoff with bad consumption outcomes will mean that the risk premium needed to tempt an investor to invest in the T-bill will be close to zero (hence Equation 11).

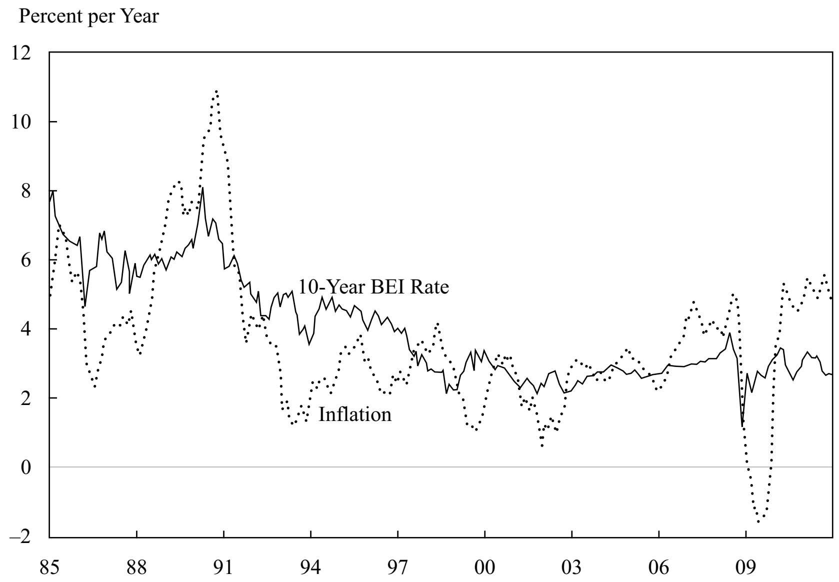
However, it is unlikely that the same level of certainty would apply, for example, to a 20-year default-free conventional government bond. For such a bond, it would seem reasonable to assume that the risk premium would be higher than that related to a one- or three-month T-bill. Note that the cash flow in Equation 10 is still certain, but only in nominal terms. Because investors will naturally have less confidence in their ability to form views about future inflation over 20 years relative to their ability to form those views over three months, the greater uncertainty about the real value of the bond's payoff will cause investors to demand a premium in compensation for this uncertainty, represented by $\pi_{t,s}$ in Equation 10.

The difference between the yield on, for example, a zero-coupon default-free nominal bond and on a zero-coupon default-free real bond of the same maturity is known as the break-even inflation (BEI) rate. It should be clear from the discussion earlier that this break-even inflation rate will incorporate the inflation expectations of investors over the investment horizon of the two bonds, $\theta_{t,s}$, plus a risk premium that will be required by investors to compensate them predominantly for uncertainty about future inflation, $\pi_{t,s}$. Although the evolution of real zero-coupon default-free yields over time should be driven mainly by the inter-temporal rate of substitution, the evolution of their nominal equivalents will, in addition, be driven by changing expectations about inflation and changing perceptions about the uncertainty of the future inflation environment. We can see this evolution by plotting the constant maturity zero-coupon break-even inflation rates over time.

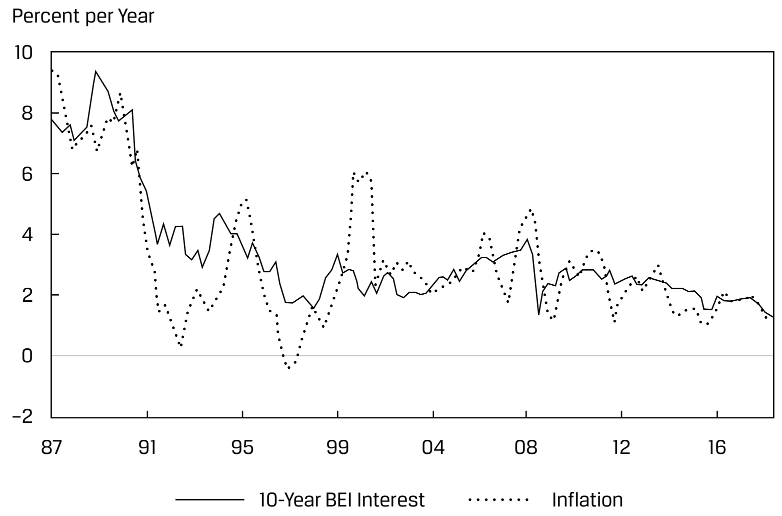
Panels A, B, and C of Exhibit 7 show the 10-year break-even inflation rates derived from three government bond markets in developed economies where index-linked government bonds have been available for some time now—Australia, the United Kingdom, and the United States—along with the respective inflation rates of each economy. The UK and Australian data, which are available for longer historical periods, show the gradual decline in break-even inflation rates since the mid-1980s. This decline was probably driven by the changing inflation environment in these economies. Between 1985 and 1990, inflation averaged approximately 6.0% and 7.5% in the United Kingdom and Australia, respectively. Between 2000 and 2011, having fallen steadily during the 1990s, inflation averaged 3.0% and 3.2% in the United Kingdom and Australia, respectively. Ten-year break-even rates in the United States were only available starting in 1997, a period when US inflation was relatively low and stable. Panel D of Exhibit 7 highlights the impact of the liquidity and credit crisis of 2008–2009 on break-even rates for a range of economies. It shows that for all of these developed economies, 10-year break-even inflation rates fell in response to the weaker global economic environment and weaker inflationary backdrop. The weaker inflationary pressure arises from the lower demand for resources in an economic downturn, so that cost and prices do not rise as fast. For example, 10-year Italian break-even inflation rates fell from 2.3% to 0.8%, reflecting the effect of the eurozone crisis on the Italian economy at that time.

Exhibit 7: Break-Even Inflation Rates and Inflation

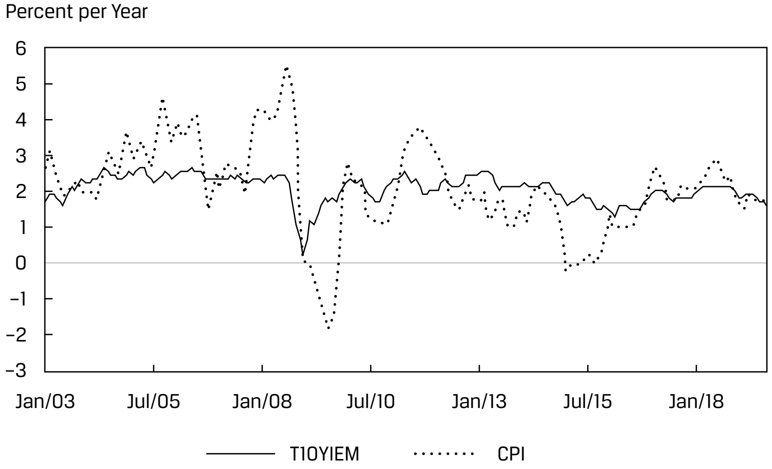
United Kingdom



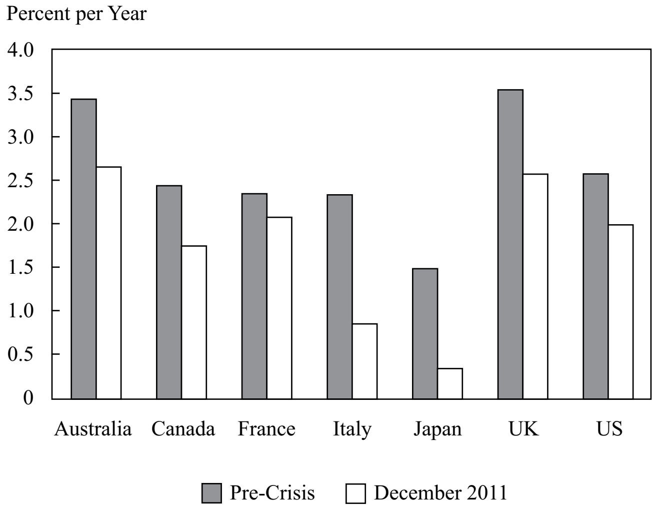
Australia



United States



Changes in BEI Rates



Sources: Based on data from Bloomberg and the authors' calculations.

Being able to measure financial market expectations of future inflation is of great value to central banks. Break-even inflation rates provide an independent view about future inflation that can be compared with the judgment of the central bank; although, of course, this judgment can be interdependent. However, it is important to remember that break-even inflation rates are not simply the markets' best guess of future inflation over the relevant investment horizon. Break-even inflation rates will also include a risk premium to compensate investors for their uncertainty largely about future inflation and, therefore, the uncertainty about the quantity of goods and services that they will be able to consume in the future.

12 THE DEFAULT-FREE YIELD CURVE AND THE BUSINESS CYCLE

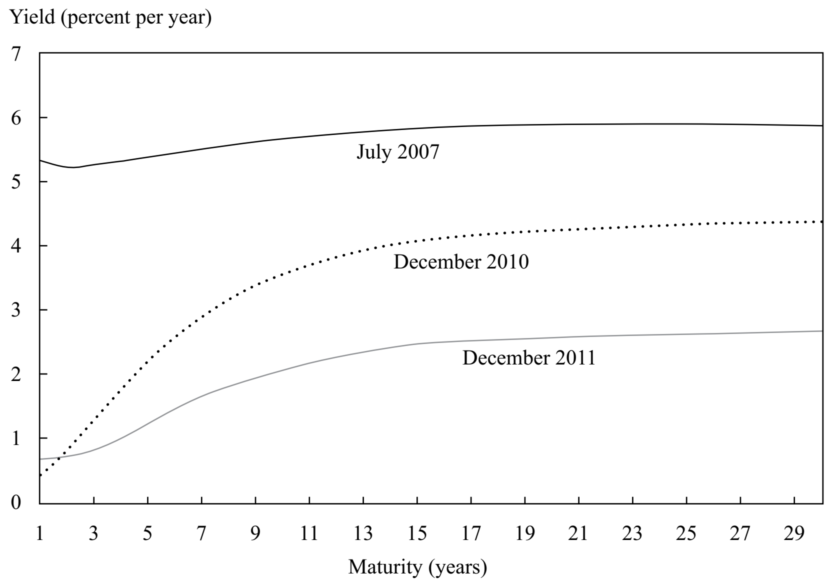
☐ explain how the phase of the business cycle affects policy and short-term interest rates, the slope of the term structure of interest rates, and the relative performance of bonds of differing maturities

So far we have discussed the fundamental pricing relationship for default-free real and nominal bonds and short-term nominal interest rates. We now elaborate on these relationships over different investment horizons. We have already indicated that the maturity of a bond will have an impact on the way that investors price it. We now focus on this relationship more specifically. But first consider Panel A in Exhibit 8, which shows the US zero-coupon Treasury curve on three different dates. From July 2007 (just prior to the wider financial crisis) to the end of 2011, the US Treasury curve shifted down by between 3 and 4 percentage points and also became steeper. The short end of the curve was clearly influenced by the reduction in the Fed's policy rate over this period, which fell from 5.25% to virtually 0%. Panel B shows that there was a similar decline in the short end of the gilt curve, as the UK central bank gradually cut its policy rate from 5.75% to 0.50% in response to the same crisis.

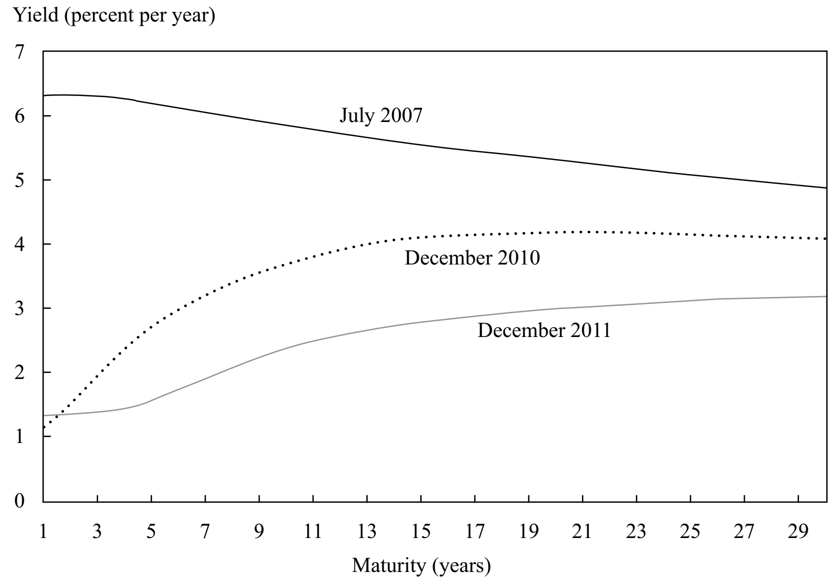
Panel A shows that the Treasury curve was upward sloping on each of these dates. Panel B shows that by the end of 2011 the UK government and US government curves looked very similar, but what is interesting is that the UK government curve was downward sloping in July 2007. This slope meant that the UK government could borrow 1-year money at 6.25% but 30-year money at 4.8%. In fact, on the same date, the UK government could borrow 50-year money at just over 3.0%. What economic factors could explain not only the fall in Treasury and gilt yields, as well as those elsewhere in the developed world, over this period but also the very negative slope of the gilt curve in the summer of 2007?

Exhibit 8: US and UK Government Bond Yields and Break-Even Inflation Rates for July 2007, December 2010, and December 2011

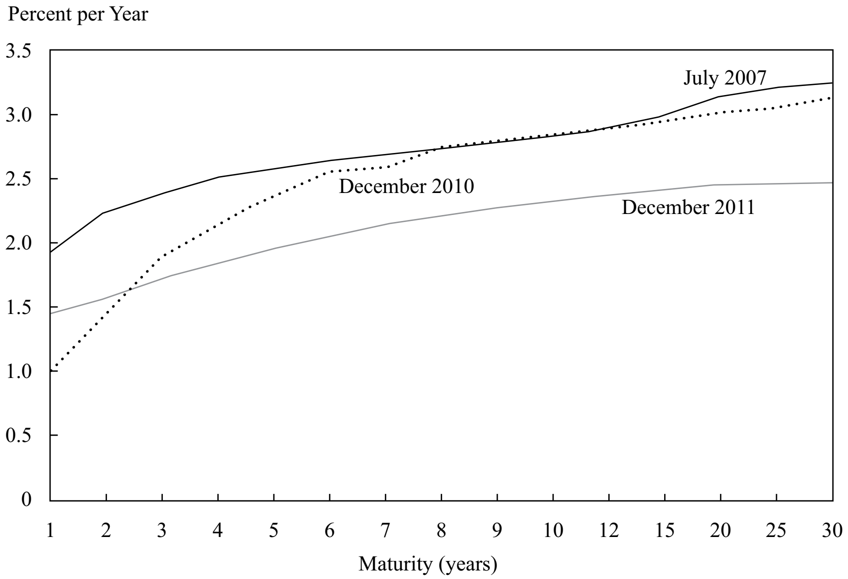
US Treasury Curve



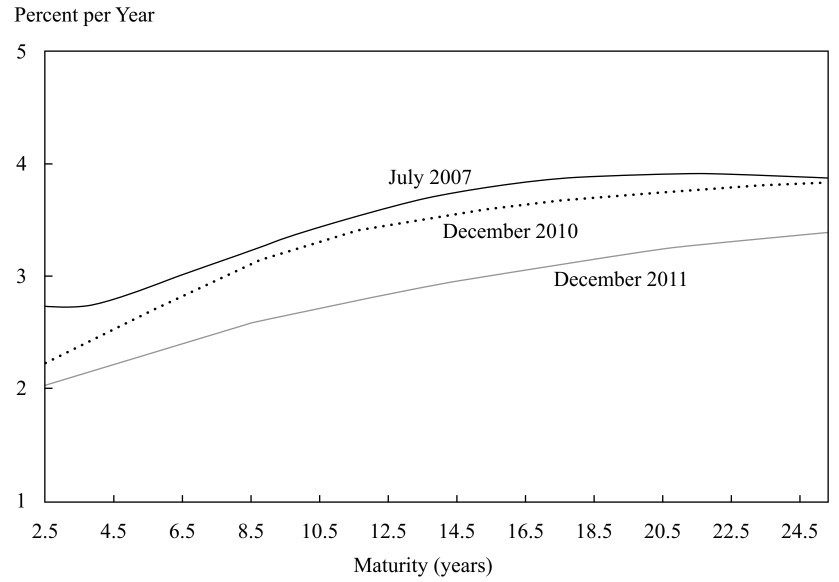
UK Gilt Curve



US BEI Rates



UK BEI Rates



Sources: Based on data from the Bank of England and Thomson Reuters.

Panels A and B of Exhibit 8 show how the shape of the default-free yield curves shifted in response to the financial crisis precipitated by the collapse of Lehman Brothers. Panels C and D of the exhibit show how break-even inflation rates shifted over the same period. When there is a significant shift in the yield curve, it is often very informative to break down the change into the real and inflationary components. Over this period, the US and UK break-even curves both shifted by just less than a full percentage point across the entire maturity spectrum. This fact could be taken to indicate that as the global recession gathered pace, US and UK government bond investors gradually "priced in" lower and lower levels of future inflation. But crucially, Panels A and B show that the nominal curves shifted down by more than 1 percentage point. For example, the US curve fell by as much as 3 percentage points. These facts suggest that market participants saw the financial crisis as potentially having a bigger impact on economic growth than on inflation. Indeed, by December 2011 the US and UK break-even inflation curves were both upward sloping.

This dissection of default-free yield curves and the interpretations that are often made by analysts based on the relative movements of the real and break-even components can be very informative, but the analysis presupposes either that there is no risk premium embedded in investors' return expectations or that any risk premium is constant over time. But the risk premium is unlikely to be zero or constant over time.

EXAMPLE 12

Level, Slope, and Curvature of the Yield Curve

1. The yield curves shown in Exhibit 8 all have three distinct characteristics. These characteristics are referred to as *level*, which indicates whether rates are high or low, on average; *slope*, which is an indicator of the steepness of the curve, or how quickly or slowly rates change with maturity; and *curvature*, an indicator of how much the curve is different from a straight line. These characteristics were first noted in rates by Steeley in 1990 and Litterman and Scheinkman in 1991. All three components can change over time. We can calculate how much of the change in any yield curve is attributable to each of these factors over time. We have performed this calculation

for the UK and US yield curves using data with maturities ranging from 3 months to 10 years, spanning the period from January 1999 to January 2014. The principal components analysis technique was used to perform the analysis. These results are shown in Exhibit 9. The majority of movement of the US (92.7%) and UK (95.2%) yield curves is in the level—that is, shifts up and down in the yield curve over time. The slope component accounts for a much smaller proportion of the change over time—6.9% for the US curve and 4.5% for the UK curve. Curvature changes account for only 0.3% in each case (note that we have rounded these figures to the nearest decimal point for convenience). Taken together then, changes in these three factors explain the vast majority of changes to these two yield curves over time. These results are typical for government bond yield curves in developed economies.

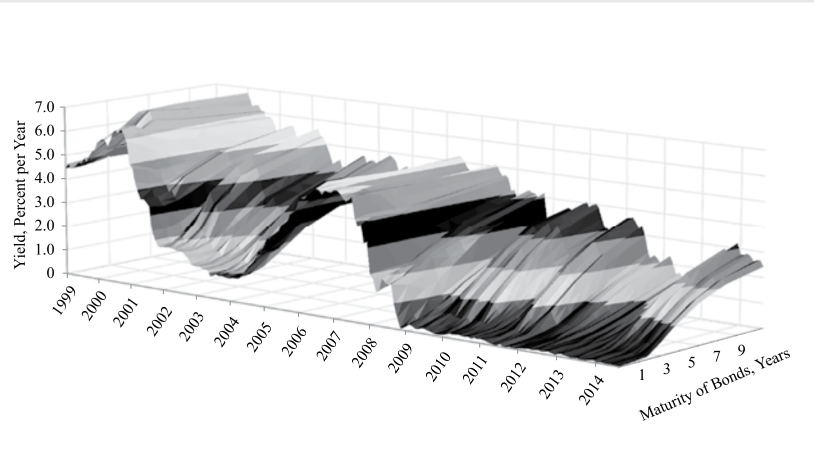
Exhibit 9: Percentage of Yield Curve Movement Explained by Three Components

	Level	Slope	Curvature
United States	92.7%	6.9%	0.3%
United Kingdom	95.2%	4.5%	0.3%

Sources: Based on data from Bloomberg and the authors' calculations.

Exhibit 10 presents another way of viewing the dynamics of the US yield curve for 15 years starting in 1999. The chart shows the way in which the level, slope, and curvature of the US Treasury curve has changed over time.

Exhibit 10: The US Yield Curve, 1999–2014



We would expect the level of economic activity to influence yield curve levels. We would also expect that views of future inflation will determine the level of these yield curves because these are nominally dominated bonds. The slope of the yield curve will be influenced by the magnitude of the risk premium in Equation 5 between the price of the bond and the inter-temporal rate of substitution over the investor's time horizon. A positive slope would be a reflection of this risk premium. However, it is not the only variable that affects the slope of the yield curve. The policy rate of the central bank is set based on Taylor rule–like considerations (that is, consideration of the components that make up the rule) so that short rates will tend to be lower during recessions because central banks tend to lower their policy rate in these times. (Note that no suggestion is made that central banks slavishly follow the Taylor rule; the rule just neatly encapsulates two of the key macroeconomic considerations that go into the process of setting the interest rate.) But the impact of monetary policy on longer-term rates will not be as strong because the central bank will usually be expected to bring short-term rates back to normal as the recession recedes and the risk-free rates will increase as economic growth recovers. Thus, the slope of the yield curve will increase during the recession. Finally, if investors anticipate that policy rates as well as short-term risk-free rates will revert back to normal as the recession recedes, then the yield curve will become steeper for the short-term maturities but flatter for the long-term maturities so that the curvature can increase as well. As a result, the shape of the yield curve and its three factors can provide valuable information for both central banks and investors.

13 THE SLOPE OF THE YIELD CURVE AND THE TERM SPREAD

☐ explain how the phase of the business cycle affects policy and short-term interest rates, the slope of the term structure of interest rates, and the relative performance of bonds of differing maturities

The required return on future default-risk-free cash flow was explained as consisting of a real interest rate, a premium for expected inflation, and a risk premium demanded by risk-averse investors for the uncertainty about what inflation will actually be (see Equation 10). Thus, referring to government yield curves, expectations of increasing or decreasing short-term interest rates might be connected to expectations related to future inflation rates and/or the maturity structure of inflation risk premiums.

Expectations of declining short-term interest rates can explain the downward-sloping UK gilt curve in the summer of 2007. If bond market participants expect interest rates to decline, then reinvestment of the principal amounts of maturing short-term bonds at declining interest rates would offset the initial yield advantage of the shorter-dated bonds. These expectations caused the United Kingdom's yield curve to be downward sloping or inverted.

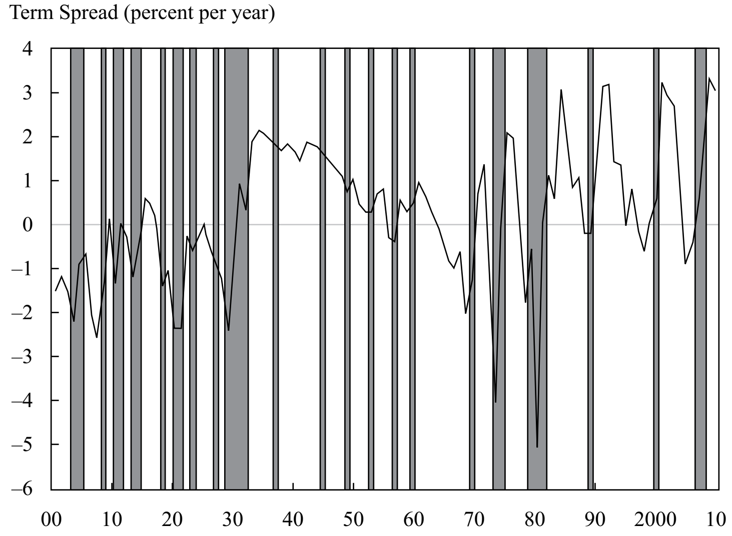
Thus, the variation in short rates over time—in particular, the central bank's policy rate—can influence the shape of the yield curve. These short rates are, in turn, driven by the positive relationship between the real rate of interest that balances investment and saving decisions over time and by the level of and volatility in GDP growth, as well as by the variation of the rate of inflation around the central bank's target, or preferred, level.

The Term Spread and the Business Cycle

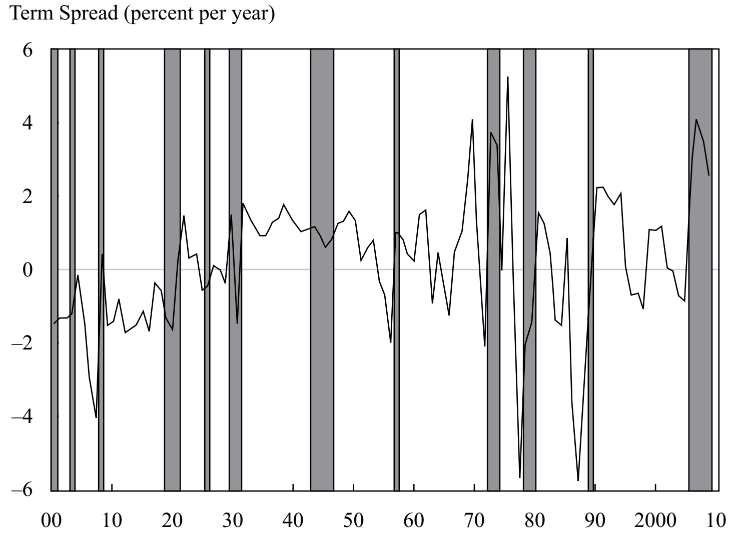
Exhibit 11 shows the time variation in the slope of the US and UK government yield curves since 1900. In both cases, the slopes have been calculated as the difference (spread) between a long-dated government bond and the yield on an equivalent one-year bond.

Exhibit 11: US and UK Government Yield Curve Spreads

US Treasury



UK Gilt



Note: Shaded areas indicate recessions.

Sources: Based on data from NBER and the Bank of England.

In both markets, there are times when the curves were very steep. For example, in the mid-1970s the steep slope implied expectations of a sharp increase in interest rates. This was a time when both inflation and inflation expectations were high, following the first oil shock of 1973.

But there are times when both curves are steeply inverted—for example, in 1979–1980. The inverted curves in these times implied an expectation of sharply falling inflation and future interest rates. In both economies, the nominal policy rates were extremely high: Policy rates peaked at 17.5% in the United States in December 1980 and at 17% in the United Kingdom in November 1979. During this period, the Fed chairman, Paul Volcker, and Margaret Thatcher had raised policy rates in their respective economies in response to the second oil shock of 1979 and administrations on both sides of the Atlantic came to the conclusion that the defeat of inflation should be the number one policy objective. The inverted curves in both markets suggest that investors expected rates to come down once the causes of high current inflation had been removed. Generally speaking, Exhibit 11 also reveals that a recession is often preceded by a flattening, or even an inversion, in the yield curve. In general, the late stages of a business expansion are often characterized by a peak in inflation and thus relatively high short-term interest rates. If longer-maturity yields reflect lower inflation rates and diminished business credit demand, the yield curve would tend to flatten or invert. An inverted yield curve, in particular, is often read as being a predictor of recession.

EXAMPLE 13

Interest Rates, the Yield Curve, and the Business Cycle

1. What financial instrument is best suited to the study of the relationship of real interest rates with the business cycle?

   A. Default-free nominal bonds

   B. Investment-grade corporate bonds

   C. Default-free inflation-indexed bonds

Solution:

C is correct. These bonds' prices are sensitive to changes in real interest rates because the payments are adjusted for changes in the price of goods.

2. Suppose investors forecast an unanticipated increase in real GDP growth and the volatility of GDP growth for a particular country. The effect of such a forecast would be for the coupon payments of an inflation-indexed bond issued by the government of the country:

   A. to rise.

   B. to fall.

   C. to be indeterminate.

Solution:

A is correct. The coupon payments would be expected to increase, reflecting an increase in the real interest rate.

3. The yield spread between non-inflation-adjusted and inflation-indexed bonds of the same maturity is affected by:

   A. a risk premium for future inflation uncertainty only.

   B. investors' inflation expectations over the remaining maturity of the bonds.

   C. both a risk premium for future inflation uncertainty and investors' inflation expectations over the remaining maturity of the bonds.

Solution:

C is correct. The difference between the yield on a zero-coupon default-free nominal bond (such as a US government STRIP) and on a zero-coupon default-free real bond of the same maturity (such as a US government TIPS) is called the break-even inflation rate. The break-even inflation rate should incorporate investors' inflation expectations over the remaining maturity plus a risk premium for uncertainty about future inflation, as in Equation 10.

4. State an economic reason why inverted yield may predict a recession.

Solution:

The late stages of business cycles are often characterized by relatively high inflation and high short-term interest rates. To the extent that longer-term yields reflect expectations of declining inflation and a slackening in demand for credit, the yield curve would be expected to flatten or invert.

EVIDENCE ON RISK PREMIUMS FOR DEFAULT-FREE BONDS

14

☐ explain how the phase of the business cycle affects policy and short-term interest rates, the slope of the term structure of interest rates, and the relative performance of bonds of differing maturities

If, as seems likely, most investors want to be compensated for taking on risk, then the yield curve, as well as containing information about the interest rate expectations of investors, will also embody a risk premium.

We have already explained why investors would value investments that paid off more in bad times relative to those investments that paid off less in these times or produced negative returns. This preference tends to drive the expected return down and the price of these favored investments up relative to those with prices that are more positively correlated with bad times. The average slopes of the US and UK government curves from 1900 to 2011 were 0.24% and 0.14%, respectively; in the post-1945 period, they were 0.50% and 0.40%, respectively. This difference suggests that, on average, investors have been willing to pay a premium for shorter-dated US and UK government bonds, which, in turn, means that longer-dated bonds may not be such a good hedge against economic bad times. One interpretation of an upward-sloping yield curve is that short-dated bonds are less positively (or more negatively) correlated with bad times than are long-dated bonds.

Exhibit 12 presents information on the relationship of government bonds with a range of maturities from a selection of countries. Panel A shows that the average yield differences (longer minus shorter) between different bond maturities, with one exception, are all positive. This fact suggests that the bond risk premium generally rises with maturity, which is why it is often referred to as the term premium. Panel B presents the total return on these government bonds by maturity. Over the sample periods, the total returns achieved generally rise with maturity in each of the bond markets. But why have government bond investors generally been rewarded for holding longer-dated government bonds relative to shorter-dated bonds?

In Panel C, we present the correlation between (1) the total return on bonds with various maturities and (2) the economic growth of the relevant economy. One thing to notice is that the correlations are predominantly negative. This fact suggests that government bonds in these markets tend to pay off in bad times, which means that investors are willing to pay a relatively high price for them. Therefore, investors should be willing to accept a relatively low return from government bonds because they are at least a partial hedge against "bad" consumption outcomes.

Exhibit 12: Government Bond Spreads, Total Returns, and GDP Growth Correlations for Four Markets

Panel A: Spreads

		5 vs. 2	10 vs. 5	30 vs. 10
Canada		0.40	0.40	
France		0.46	0.52	0.60
UK		0.27	0.26	−0.05
US		0.55	0.44	0.39

Panel B: Total Returns

	2	5	10	30
Canada	5.61%	6.70%	8.08%	9.60%
France	4.44%	6.09%	7.80%	10.24%
UK	7.14%	7.82%	10.12%	11.03%
US	5.77%	6.91%	7.72%	9.07%

Panel C: Correlation with GDP

	2	5	10	30
Canada	−4.93%	−5.12%	−0.70%	2.75%
France	4.71%	−9.14%	−10.27%	−2.13%
UK	−4.66%	−2.74%	−5.11%	0.79%
US	−7.30%	−14.33%	−12.31%	−10.35%

Note: The sample period for the United Kingdom and the United States is January 1980 to December 2018. For France and Canada, it is January 1985 to December 2018.
Sources: Thomson Reuters and authors' calculations.

The results shown in Exhibit 12 suggest that government bond risk premiums

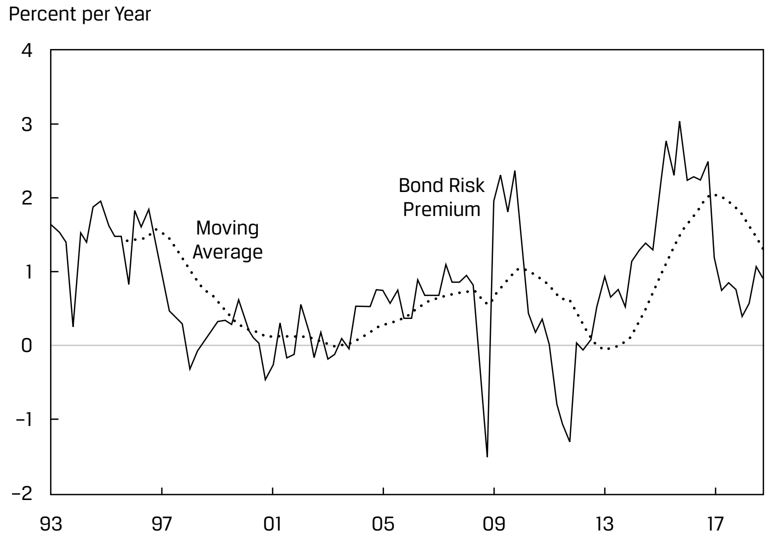
- are positive,
- are probably related to the consumption hedging benefits of government bonds, and
- are positively related to bond maturity, which means that the "normal" shape for the yield curve is upward sloping.

The last point also helps to explain why the US Treasury curve was generally upward sloping between the summer of 2007 and the end of 2011: A significant portion of the slope was probably related to the existence of a positive risk premium on US Treasuries that increased with maturity.

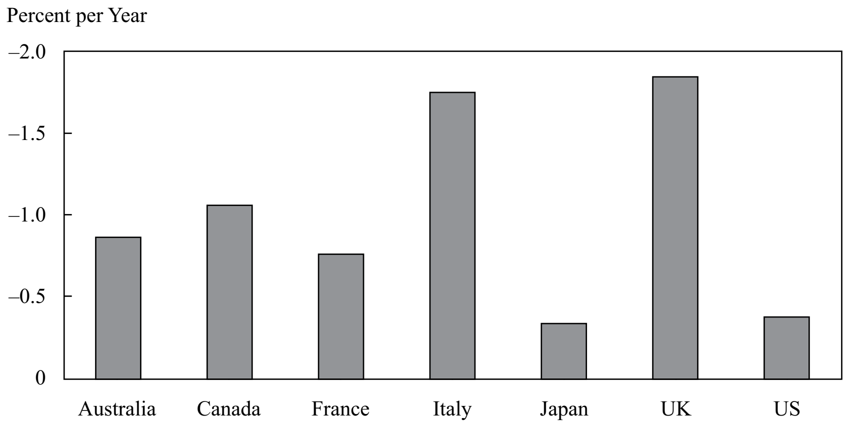
However, bond risk premiums ($\pi_{t,s}$), like other risk premiums, will not be constant over time. In times of economic uncertainty, investors will tend to more highly value assets that pay off in bad times—government bonds—which will force their prices up as the risk premium demanded falls. Unfortunately, it is impossible to say how big or small this premium really is or should be or how it evolves with the business cycle. But we can get some idea by performing the following experiment. Suppose we can assume that the real return on an index-linked government bond is a good proxy for the real rate of interest that balances savings and investment in an economy over time. If we subtract the yield on an index-linked government bond from the yield on a conventional government bond with a similar maturity and then subtract from this amount a survey-based measure of inflation expectations, what is left over is the bond risk premium. That is, the extra yield investors require for holding a conventional government bond over and above the real required return and the return in compensation for expected inflation.

Exhibit 13: Bond Risk Premiums (BRPs)

UK Government BRP



Change in Government BRP, 2007–2011



Note: The period 1993–1997 in Panel A was before the inflation-targeting regime.

Panel A of Exhibit 13 presents a calculation of this kind for the UK government bond market. There are a number of points worth noting with respect to this particular representation of the bond risk premium. First, it is certainly imperfect because it shows that the bond risk premium is negative at times. A negative bond risk premium implies that investors are willing to pay for assuming risk, which is inconsistent with risk aversion. Second, putting aside these negative values for the moment, we can see that the bond risk premium varies over time. Between 1993 and mid-1997, the risk premium on UK government bonds averages just less than 1.5%; from mid-1997 to 2005, it averages around 0.25%; and from 2013 to 2018, it averages around 1.42%.

Panel B in Exhibit 13 is based on similar estimates for the bond risk premiums over the global financial crisis for a range of markets and shows the change in the bond risk premiums from July 2007 to December 2011. For all the government bond markets for which we could make these calculations, the bond risk premium fell—in some cases, quite substantially. This fact implies that investors placed greater value on the consumption-hedging properties of government bonds as a result of the financial crisis. Other things being equal, this implies that they were willing to pay a higher price for these bonds compared with the price they were willing to pay before the crisis. This also implies that bond risk premiums will tend to rise in times when investors place less value on the consumption hedging properties of government bonds.

15 OTHER FACTORS

☐ explain how the phase of the business cycle affects policy and short-term interest rates, the slope of the term structure of interest rates, and the relative performance of bonds of differing maturities

In practice, the shape of the yield curve and the relative performance of bonds with different maturities over the business cycle depend on a complex mixture of interest rate expectations and risk premium considerations. For example, a downward-sloping curve is probably largely the result of investor expectations of future declines in interest rates. The drivers of an upward-sloping curve are more ambiguous. For example, the existence of bond risk premiums that are positively related to maturity means that an upward-sloping curve may not embody expectations of future rate increases. Conversely, it could imply a combination of expected rate increases and risk premiums or even expected rate cuts that are more than offset by the existence of positive risk premiums.

We have to acknowledge that there will also be times when other factors play a part in shaping the yield curve. Developed-economy government bonds are technically default-free because these governments can, in principle, always print cash to meet the promised payments. In this sense, they are very special financial instruments. These markets can also be influenced by supply and demand factors that seem to move yields in ways that do not appear to be consistent with the business cycle. Consider the following examples:

- In the late 1990s, fiscal surpluses in the United States led some investors to take the view that the supply of Treasuries would shrink as the US government paid back its debts (leading to the Treasury scare of 2000). This perceived reduction in future supply was said to have been responsible for the decline in yields as investors bought up these bonds in anticipation of their future scarcity.
- In the early 2000s, Treasury yields were apparently being pushed down by Asian central banks that were using their growing trade surpluses to purchase US Treasuries.
- Regulatory factors can also play an important role in determining government bond yields. This influence is particularly clear in the gilt market. In 1997, the UK government passed legislation that effectively compelled UK pension funds to buy long-dated gilts. This legislation appears to have been one of the main drivers of the inversion of the long end of the United Kingdom's yield curve (see Exhibit 8, Panel B). Since that time, new accounting rules for pensions—FRS17, followed by IAS19—forced UK pension schemes to increase their demand for long-dated UK government bonds further still. These actions created a vicious circle because the new accounting rules required the schemes to discount their liabilities using the long-dated yields as the discount rate. Consequently, UK pension funds bought long-dated bonds, forcing their yields down. The decline in yields caused the present value of pension liabilities to rise relative to the value of

scheme assets, creating deficits. To achieve a better match between assets and liabilities, schemes tried to buy more long-dated gilts, causing their yields to fall further and liabilities to rise further. A similar phenomenon affected core eurozone bond and swap markets. For more than three years, the yield spread between 10- and 30-year Dutch government bonds (30 minus 10) was negative. It was argued that the negative spread was a direct result of the hedging activities of Dutch pension funds, which had over €1 trillion of pension liabilities. The buying pressure on a government bond market that was only around 60% the size of these liabilities thus caused long-dated Dutch government bond prices to rise.

There is no doubt that supply and demand considerations along with poorly thought out regulatory or accounting rules can have an impact on government bond markets. But determining the extent of these effects is very difficult. Nevertheless, it is difficult to explain the very inverted shape of the long end of the gilt curve, which has persisted for many years now, without reference to such factors.

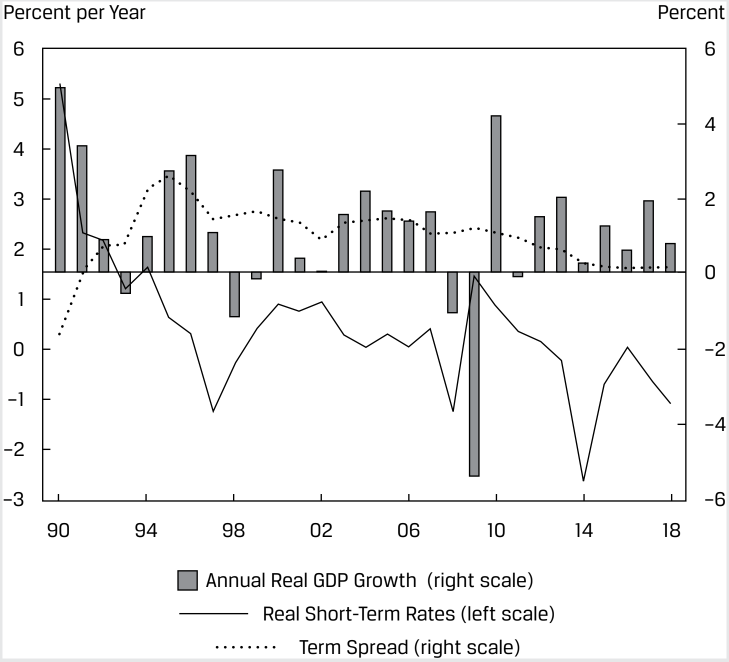
EXAMPLE 14

The Japanese Yield Curve and Business Cycle, 1990–2018

1. During the 1980s, real Japanese GDP growth averaged more than 4.60% per year. During the 1990s and 2000s, it averaged around 1.5%. The catalyst for the decline in growth was the collapse of Japan's property bubble and the stock crash in the early 1990s. Exhibit 14 shows Japan's real annual GDP growth over the post-bubble period, which incorporates the global financial crisis of 2008–2009. The exhibit also shows how real short-term policy rates fell steadily from a level of 4%–5% in 1990, finally becoming negative by 1997. Since that time, real rates have generally been negative as Japan's cen-

tral bank, the Bank of Japan, has tried to stimulate its economy by cutting its nominal policy rate to (near) zero.

Exhibit 14: Japan's GDP Growth, Real Short-Term Rates, and Term Spread, 1990–2018



Sources: Based on data from Thomson Reuters and the authors' calculations.

Exhibit 14 also shows the term spread over this period—that is, the slope of the yield curve (created by subtracting the policy rate from 10-year Japanese government bonds). At the start of the period, as the Bank of Japan cut its policy rate from just over 8.0% at the end of 1990 to 0.45% by the end of 1995, the curve steepened—from −1.63% at the end of 1990 to 2.59% by the end of 1995. The sharp steepening of the curve in response to the cuts in the policy rate arguably indicated the market's view that the policy stimulus would work; that is, the yields on longer-dated Japanese government bonds embedded some expectation of positive growth and inflation in the future. However, since the mid-1990s, despite cutting its policy rate to zero and enacting programs of quantitative easing, the term spread has fallen steadily. By the end of 2018, it was virtually 0.00%. The weak economic growth and inflation environment in Japan since the collapse of the property bubble is reflected in the fall and flattening of the Japanese yield curve.

16 CREDIT PREMIUMS AND THE BUSINESS CYCLE

☐ explain how the phase of the business cycle affects credit spreads and the performance of credit-sensitive fixed-income instruments

Earlier we discussed the economic drivers of what we have referred to as default-free interest rates and bond yields. But the financial crisis has caused many to question what "default-free" really means. The bonds issued by many European governments, including those issued by France and Italy as well as those issued by the Greek, Portuguese, Irish, Belgian, and Spanish governments, were all thought to be default-free before the euro financial crisis from 2010 to 2012. Even the default-free status of US Treasuries has now been questioned by both investors and credit rating agencies. Any bond that is perceived to be default-free will, by definition, not have to compensate investors for taking on default, or credit, risk. However, for any corporate or government bond that embodies the non-zero probability that the issuer may default on its obligations, bondholders will demand a risk premium, referred to as the credit premium.

In Equation 1, we emphasized that the discount rate on the cash flows of financial assets will normally include a risk premium, which we defined generically as $\rho_{t,s}^{i}$. When we considered the pricing formula for default-free government bonds, we emphasized that the risk premium attached to these bonds, $\pi_{t,s}$, is largely a function of uncertainty about future inflation and that this uncertainty is likely to be greater for longer-dated bonds relative to shorter-dated bonds (recall that this risk premium, $\pi_{t,s}$, is distinct from the addition to return that investors require on the basis of their expectations of inflation over the investment horizon, $\theta_{t,s}$, per Equation 1). But it is important to emphasize that these factors will play a role in the pricing of bonds that embody credit risk, too. In other words, the evolution of $\pi_{t,s}$ plays a role in determining the price of credit risky bonds too. We thus need to acknowledge the separate role that credit risk plays in the price of a corporate bond or, indeed, any bond with default risk. We have therefore adapted our basic pricing Equation 1 to augment the discount rate with a credit premium, $\gamma_{t,s}^{i}$, which is distinct from the inflation-based risk premium attached to default-free bonds, $\pi_{t,s}$, as shown in Equation 13.

$$P_t^i = \sum_{s=1}^{N} \frac{E_t\left[\widetilde{CF}_{t+s}^{i}\right]}{\left(1 + l_{t,s} + \theta_{t,s} + \pi_{t,s} + \gamma_{t,s}^{i}\right)^{s}} \tag{13}$$

Notice that Equation 13 acknowledges the uncertain nature of the cash flows, $CF_{t,s}$, on credit risky bonds; although the schedule of these payments is known, the existence of credit risk means that there will be uncertainty about whether they will be paid as scheduled. In the event of a default, the amount that the bond investor receives will depend on the recovery rate, which will also be an unknown quantity. The risk premium demanded by investors because of these uncertainties is represented by $\gamma_{t,s}^{i}$ in Equation 13.

In this section, we will focus on the credit premium, $\gamma_{t,s}^{i}$, and in particular on the relationship between the business cycle and the credit premium on corporate bonds.

EXAMPLE 15

The Credit Risk Premium

1. Suppose that an analyst estimates that the real risk-free rate is 1.25%, average inflation over the next year will be 2.5%, and the premium required by investors for inflation uncertainty is 0.50%. If the analyst observes the price of a corporate bond with a face value of £100, with one full year to maturity, as being equal to £94.21, what would be the implied credit premium embedded in the bond's price for inflation uncertainty?

Solution

The (approximate) implied premium can be calculated as follows:

$$\gamma_{t,s}^{i} = 1.90\% = \frac{100}{94.21} - (1 + 1.25\% + 2.50\% + 0.50\%).$$

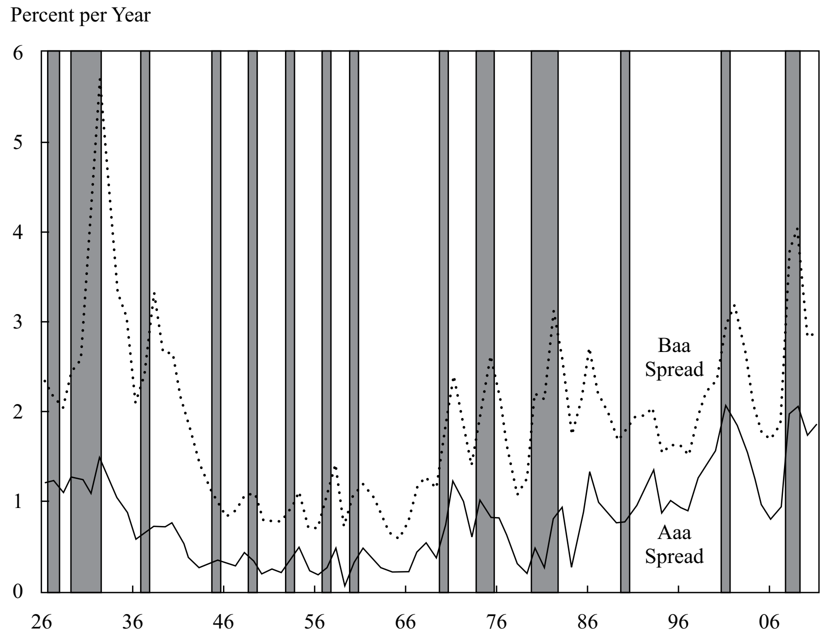
Credit Spreads and the Credit Risk Premium

The difference between the yield on a corporate bond and that on a government bond with the same currency denomination and maturity is generally referred to as the credit spread. It is demanded by investors in compensation for the additional credit risk that they bear compared with that embodied in the government bond.

As Equation 13 shows, credit risky bonds share the same risk as default-free bonds, which market participants often refer to as interest rate risk, but they also embody credit risk, $\gamma_{t,s}^{i}$. Other things being equal, a parallel shift up in the yield curve will have an almost identical proportionate impact on the prices of, say, a five-year government bond and a five-year corporate bond. And over time, again other things being equal, the interest rate component of a corporate bond will be driven by the same factors that drive government bond yields and returns. In other words, they are both subject to interest rate risk.

It is the credit risk component of a corporate bond, $\gamma_{t,s}^{i}$, and the evolution of bond spreads that will cause corporate and comparable government bond returns to diverge over time. It would seem sensible to assume that the premium demanded would tend to rise in times of economic weakness, when the probability of a corporate default and bankruptcy is highest. Exhibit 15 confirms this view. The exhibit shows a representative spread on both AAA/Aaa and BBB/Baa rated US corporate bonds over US Treasuries. (The AAA rating category in "AAA/Aaa" is the rating category used by both Standard & Poor's and Fitch Ratings; Aaa is the equivalent rating category used by Moody's Investors Service.) First, the Baa spread is always higher than the Aaa spread, reflecting the lower credit quality of Baa rated bonds relative to Aaa rated bonds. Second, the US recession periods shaded grey in the chart indicate that both low- and higher-grade corporate bond spreads do tend to rise in the lead-up to and during a recession and to decline once the economy comes out of recession.

Exhibit 15: Credit Spreads and the Business Cycle



Note: Shaded areas indicate recessions.

Source: Based on data from Moody's Investors Service.

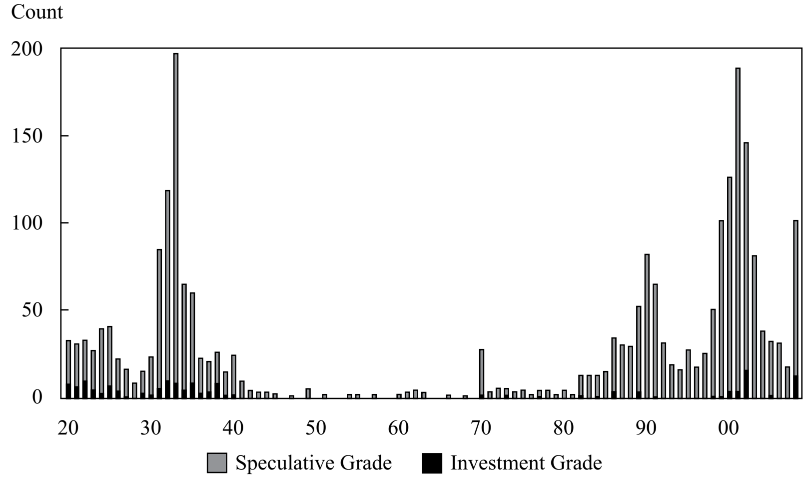
As expected, the business cycle has a profound effect on credit spreads, but what are the basic components of the credit spread? If we assume that investors are risk neutral, then they will simply demand a return (yield) on their corporate bond investments sufficient to compensate them for the possible loss that they could incur from holding a corporate bond. In turn, this expected loss will depend on the probability of default and the expected recovery rate in the event of default, as shown in Equation 14:

$$\text{Expected loss} = \text{Probability of default} \times (1 - \text{Recovery rate}). \quad (14)$$

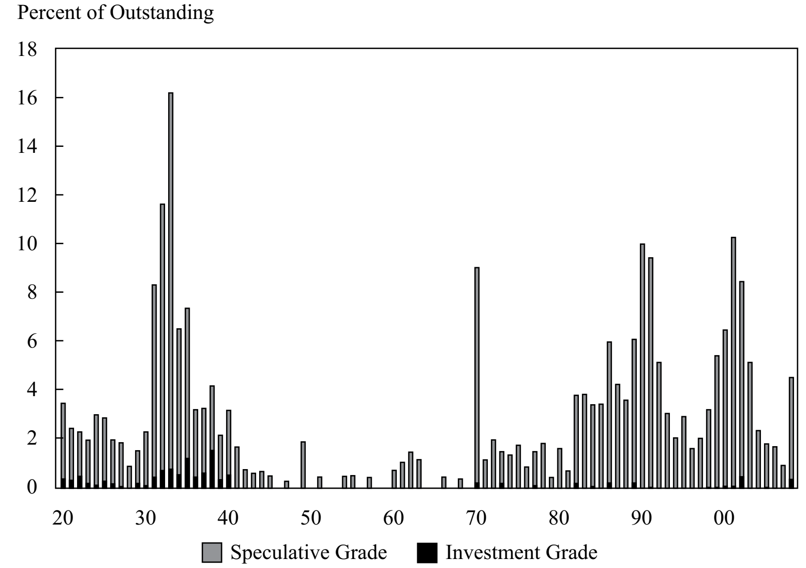
In the instance where investors are risk neutral, the expected return on, say, a 10-year government bond would be equal to the loss-adjusted expected return on a comparable 10-year corporate bond. In practice, however, investors are risk averse, so the expected return on a corporate bond will be higher than that on a comparable government bond, even if a significant amount of the credit risk can be mitigated by holding a diversified portfolio of corporate bonds. One of the main reasons why investors continue to be exposed to considerable market risk even in a well-diversified portfolio is that defaults tend to cluster around downturns in the business cycle. Panels A and B of Exhibit 16 show this quite clearly. Panel A shows the number of US corporate defaults per year since 1920, and Panel B shows annual default rates over the same period. Both charts show that there are often long periods of time when there are very few defaults. However, the US depression of the 1930s and the recessions in the 1980s, 1990s, and 2000s were all associated with relatively high default levels and rates. The historical default rates on different ratings classes are sometimes used by analysts as a proxy for the probability of default in Equation 14 for expected loss.

Exhibit 16: US Corporate Defaults, Default Rates, Recovery Rates, and Loss Rates

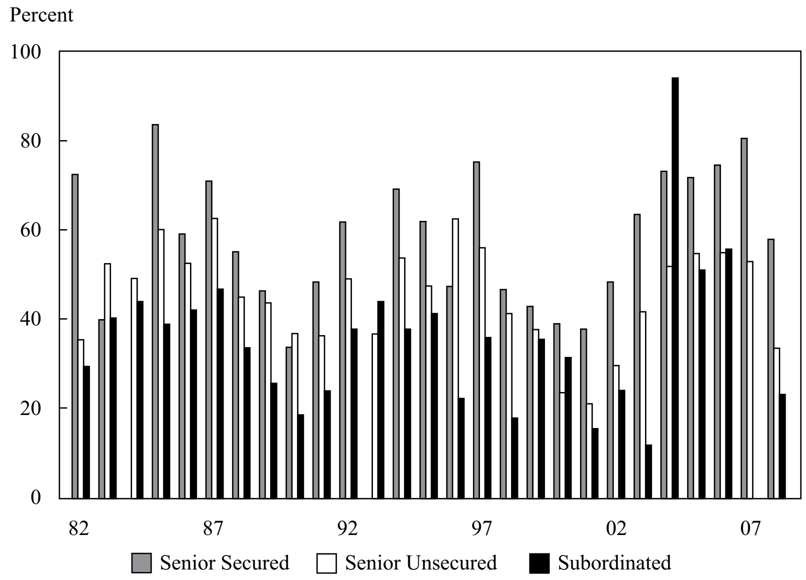
Number of Defaults



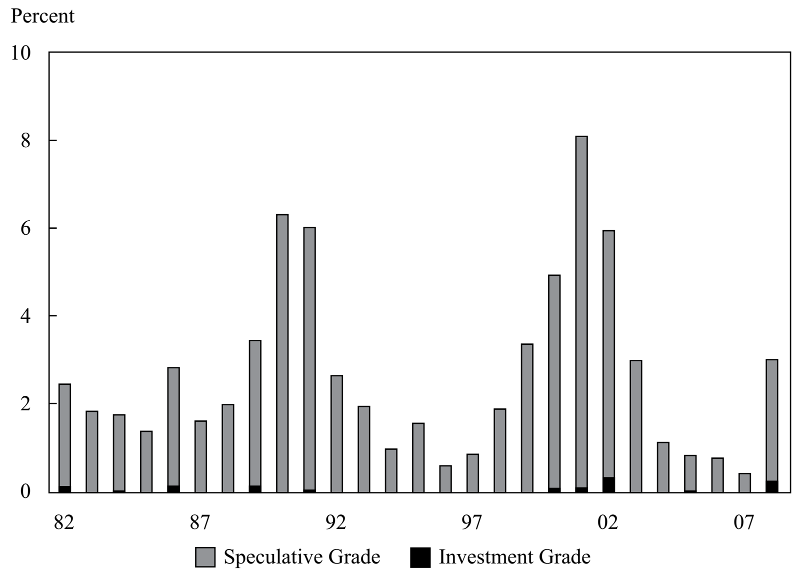
Default Rates



Recovery Rates



Loss Rates



Source: Based on data from Moody's Investors Service.

Panel C of Exhibit 16 shows the evolution of recovery rates on three types of corporate bonds: senior secured, senior unsecured, and subordinated. Senior secured debt, as the name suggests, is secured by a lien or other claim against some or all of the company's assets, whereas senior unsecured debt has no explicit claim to the company's assets in the event of bankruptcy. This explains why recovery rates are generally higher for secured as opposed to unsecured debt holders. Subordinated debt holders, as the name suggests, have an inferior claim on the company's assets compared with senior debt holders, and unsurprisingly, recovery rates are often very low. The recovery rate ranking is interesting, but we can also see from Panel C that recovery rates tend to be higher when the economy is expanding and lower when it is contracting. The reason is that assets that can be sold in order to recover value for bond holders are likely to fetch a higher price in a buoyant economic environment than in a stagnant one. Finally, Panel D shows the loss rates on US corporate debt from

1982 to 2008. These loss rates are the net result of the defaults and recovery rates over time. Unsurprisingly, these loss rates are counter-cyclical with regard to the business cycle, meaning that they tend to rise as economic activity declines.

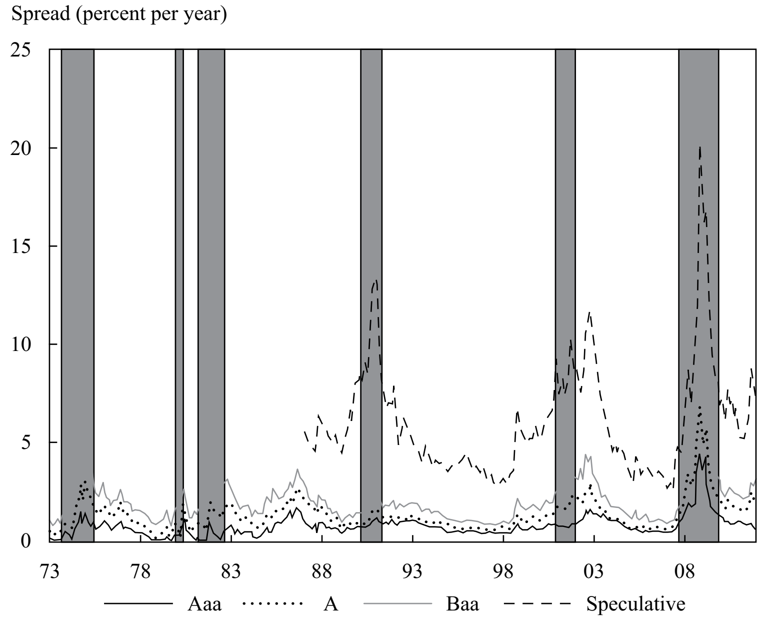
17 INDUSTRY- AND COMPANY-SPECIFIC CREDIT QUALITY

☐ explain how the characteristics of the markets for a company's products affect the company's credit quality

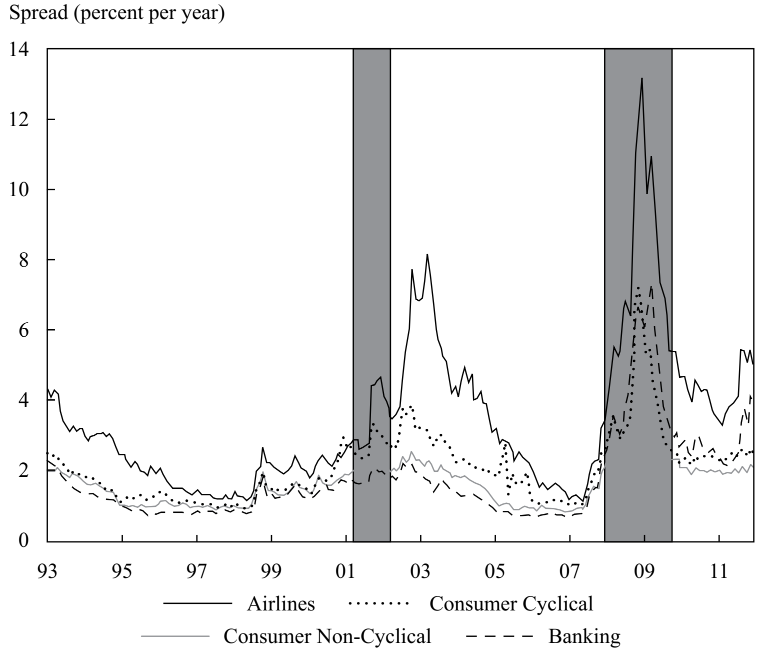
Although spreads will evolve with the business cycle, Exhibit 17 illustrates that spreads between corporate bond sectors with different ratings will often have very different sensitivities to the business cycle. Panel A presents a shorter but finer picture of the relative performance of US corporate bonds by Moody's rating category. The graph shows that when spreads are narrowing relative to government bonds, the spreads between higher- and lower-rated bond categories also narrow. In these times, although corporate bonds will generally outperform government bonds, lower-rated corporate bonds will tend to outperform higher-rated bonds. The converse is true as spreads widen, a phenomenon that is illustrated most graphically following the collapse of Lehman Brothers in 2008. The spread on speculative, or high-yield, debt rose from a pre-Lehman Brothers collapse low of around 2.8% to a peak of just more than 20.0%. Over the same period, Baa rated debt spreads rose from around 1.1% to 8.5% and Aaa corporate bond spreads rose from 0.6% to 4.5%.

Exhibit 17: US Credit Spreads and the Business Cycle

Moody's Rating



Industrial Sector



Note: Shaded areas indicate recessions.

Sources: Based on data from Thomson Reuters and the authors' calculations.

Panel B in Exhibit 17 illustrates another determinant of credit spreads: industrial sector. Analysis by industrial sector addresses the question of how the type of goods and services that individual companies produce may be related to credit quality. Some industrial sectors are more sensitive to the business cycle than others. This sensitivity can be related to the types of goods and services that they sell or to the indebtedness of the companies in the sector. Panel B shows evidence of only the divergent performance of corporate bond sectors over a relatively short period, but the performance of the four sectors shown is very different in times of economic stress. First, in both of the recessions that this period covers (indicated by the shaded areas in Panel B), the spread on the consumer cyclical sector rose more dramatically than it did for corporate bonds in the consumer non-cyclical sector. For example, the spread on the consumer cyclical sector peaked at just under 4.0% in 2003, compared with around 2.5% for the consumer non-cyclical sector. Second, the graph shows how sensitive the airline sector's credit spread is to the business cycle. The sharp widening of spreads in this sector as a result of both recessions is probably also a function of the lower credit quality, on average, of companies in the airline sector. Third, the recession of the early 2000s had only a mild impact on the spreads of banks but a much larger impact in the post-Lehman period, when the sector spread peaked at nearly 7.5%. This difference highlights the fact that the last recession and crisis were first and foremost a banking crisis. But perhaps the most interesting feature is the narrowing of sector spreads in the summers of 1998 and 2007. In both of these periods, investors were content to receive virtually the same credit spread, $\gamma^{i}_{t,s}$, on airline company debt as on debt issued by companies in the consumer non-cyclical sector.

Company-Specific Factors

Corporate bond spreads will be driven over time by the business cycle, but the impact of the economic environment on spreads will depend on issuers' industrial sector and rating. When spreads widen, the spreads on bonds issued by corporations with a low credit rating and/or that are part of a cyclical sector will tend to widen the most.

Company-specific factors will also play a part in determining the difference in the yield of an individual corporate issuer and that of a government bond with the same maturity. Issuers that are profitable, have low debt interest payments, and are not heavily reliant on debt financing will tend to have a high credit rating because their ability to pay is commensurately high.

Exhibit 18 provides summary statistics on financial statements for companies across a range of Moody's rating categories. Pre-tax interest coverage is calculated by dividing total pre-tax earnings by total debt interest payments. On average, Aaa companies had $17.60 of pre-tax earnings for every $1 of interest payment to which they were committed. By contrast, on average, Baa companies had only $2.50, whereas the average B and Caa rated companies could not cover their interest payments with current-period pre-tax earnings. The ratio of free operating cash flow to total debt gives another indication of the profitability and financial flexibility of a company relative to its outstanding debt. There is again a clear deterioration in this metric as average rating quality declines. Finally, the ratio of total debt to total capital gives an idea of the overall indebtedness of a company. Together these and other ratios allow analysts and credit rating agencies to determine a company's ability to meet its debt obligations as they come due. If this ability declines relative to other issuers in their sector, then the spread demanded on their debt will rise, relative to the sector average, and their rating may be lowered by the rating agency.

Exhibit 18: Ratings and Financial Ratios

	Aaa	Aa	A	Baa	Ba	B	Caa
Pre-tax interest coverage (×)	17.6	7.6	4.1	2.5	1.5	0.9	0.7
Free operating cash flow/Total debt (%)	42.3	28	13.6	6.1	3.2	1.6	0.8
Total debt/Total capital (%)	21.9	32.7	40.3	48.8	66.2	71.5	71.2

Source: Based on data from Moody's Investors Service.

18 SOVEREIGN CREDIT RISK

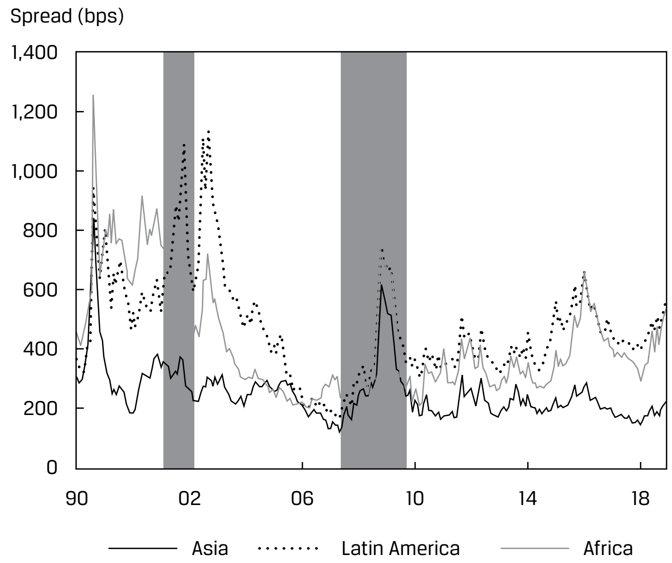
☐ explain how the phase of the business cycle affects credit spreads and the performance of credit-sensitive fixed-income instruments

So far we have discussed the risk premium demanded by investors on both default-free and corporate debt. But credit premiums have always been an important component of the expected return on bonds issued by governments in developing or emerging economies. Even though many of these governments can print money to meet their debt obligations *in extremis*, meaning that they could technically avoid defaulting on these debts, many developing-economy governments have defaulted on their debts in the past. For example, the Russian government defaulted on its debt in 1998, and many others, including Argentina, Brazil, and Mexico, have also defaulted. Such defaults are often very country specific in character, but the global economic environment, oil prices, and the evolution of global trade will often play a part in precipitating such sovereign defaults.

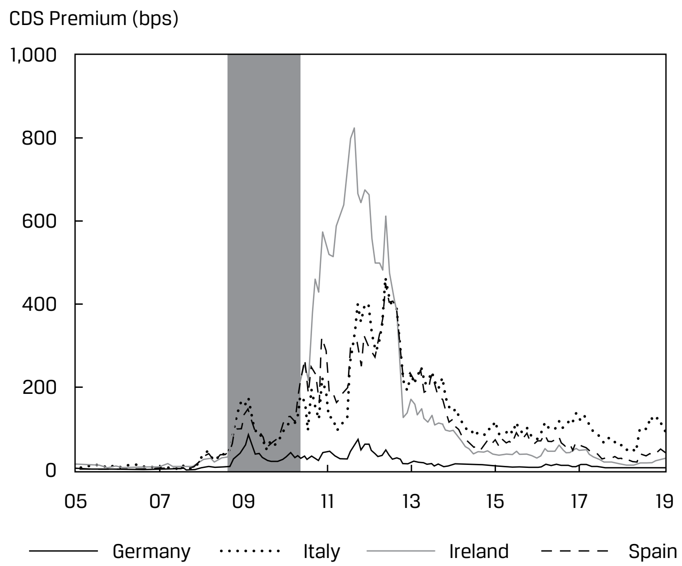
The credit risk embodied in bonds issued by governments in emerging markets is normally expressed by comparing the yields on these bonds with the yields on bonds with comparable maturity issued by the US Treasury. Panel A of Exhibit 19 shows the evolution of this spread for three emerging market bond indexes. The impact of the credit crisis is clear; spreads rose in response to the uncertain economic environment globally. But the volatility of the spreads from 1998 to 2003 is a function of the Asian financial crisis in 1997, the Russian debt crisis in 1998, and the recession in developed economies in 2001–2002 following the collapse of the high-tech bubble. What is interesting is the decline in spreads for US Treasuries up to 2007, along with the much narrower spreads between the regions at this time. Strong global economic growth between 2003 and 2007 convinced investors that they did not need such a high reward for emerging market default risk and that they did not need to differentiate much between regions.

Exhibit 19: Sovereign Credit Spreads

Emerging Markets



Developed Economies



Note: Shaded areas indicate US recessions.

Sources: Based on data from Thomson Reuters and the authors' calculations.

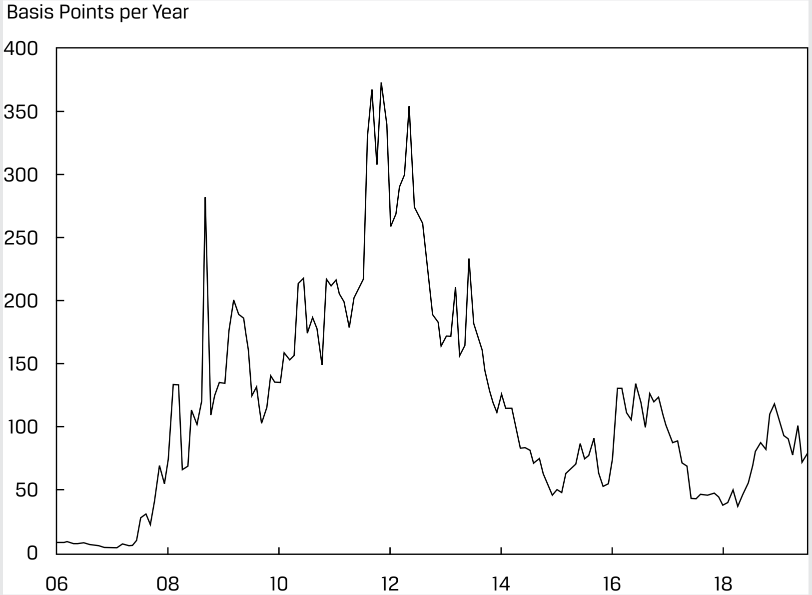
The 2008–2009 global financial crisis caused many investors to question what is meant by the term "default-free." Until that time, there was thought to be a set of developed-economy government issuers for which the likelihood of debt default was so low that it could almost be ignored. In other words, investors did not demand meaningful compensation for assuming this risk. Panel B of Exhibit 19 shows how this perception changed dramatically for the debts of a set of eurozone economies during the crisis. The chart shows the cost that an investor would have to pay for insuring themselves against a sovereign default on German, Italian, Irish, and Spanish government debt over the next five years through the purchase of credit default swaps (CDSs). For example, in January 2006 that cost was 1.8, 8.8, 2.5, and 2.8 bps, respectively. So, to insure oneself against a default on bonds issued by these governments with, say, a notional value of €10 million would have cost €1,800, €8,800, €2,500, and €2,800, respectively, per year. However, by August 2011, this insurance cost had risen to €59,830, €306,860, €825,390, and €300,610, respectively. Although the causes of this reassessment of sovereign credit risk inherent in developed-economy debt were complex, the basic reason for the increase in the credit risk premium was a reassessment by investors of these sovereign issuers' ability to pay and the likelihood that they might default. The perception of their ability to pay deteriorated dramatically as private-sector debts were absorbed onto sovereign balance sheets. And so, to some extent, the rise in this insurance cost was related to the balance sheets of these sovereign nations in much the same way that a deterioration in the quality of the balance sheet of a corporate borrower would cause its credit spread to widen.

EXAMPLE 16

The Credit Premium for the Royal Bank of Scotland

The global financial crisis had an impact on the prices of all financial assets. Exhibits 15 and 17 show the impact of the crisis on credit spreads derived from indexes, whereas Exhibit 19 demonstrates how the same crisis caused the credit spreads on some sovereign issuers to rise dramatically too. In Exhibit 20, we focus on the impact of the crisis on the CDS premium (a close proxy for the credit spread) on five-year Royal Bank of Scotland senior unsecured debt. The exhibit shows the same increase in the perception of credit risk. The eventual decline in the premium was a consequence of the UK government's nationalization of this systemically important global bank.

Exhibit 20: CDS Premium on Five-Year Royal Bank of Scotland Senior Unsecured Debt



Source: Based on data from Thomson Reuters.

Credit Premium Summary

The credit premium ($\gamma_{t,s}^{i}$) is the additional yield required by investors over and above the yield required on comparable default-free debt that investors demand for taking on credit risk. It will tend to rise and fall with the business cycle, mainly because credit risk will tend to rise as an economy turns down and to fall as an economy turns up. However, when credit spreads are generally narrowing, the rate of improvement will tend to be greater for those bonds issued by entities with a relatively weaker ability to pay. At these times too, investors seem to be less discerning among issuers with weak and strong credit credentials. But as the business cycle turns down, those issuers with a good credit rating tend to outperform those with lower ratings as the spread between low- and higher-quality issuers widens. This relationship between the economic cycle and defaults means that credit risky bonds (corporate or sovereign) tend to perform poorly in bad economic times, and because of this tendency, investors demand a credit premium.

EQUITIES AND THE EQUITY RISK PREMIUM 19

☐ explain the relationship between the consumption hedging properties of equity and the equity risk premium

Earlier we discussed the credit risk embedded in a bond that has been issued by either a corporation or government, which might not honor its promise to pay the coupons and principal payment in full and on time. Investors can thus not be certain that they will receive the future scheduled cash flows from credit risky bonds. However, when investors purchase bonds that embody credit risk, normally they at least know the proposed schedule of payments and how they are to be determined. But there are other financial instruments in which both the size and timing of the cash flows are uncertain and, indeed, where the cash flows may not materialize at all. The best example of a security that has cash flows with these characteristics is equity because the dividend payment is not promised, can rise and fall over time, and in the event that the issuing corporation becomes bankrupt, can cease altogether.

For equities, we can rewrite the generic pricing equation, Equation 1, as follows:

$$P_t^i = \sum_{s=1}^{\infty} \frac{E_t\left[\widetilde{CF}_{t+s}^i\right]}{\left(1 + l_{t,s} + \theta_{t,s} + \pi_{t,s} + \gamma_{t,s}^i + \kappa_{t,s}^i\right)^s} \quad (15)$$

Notice that this equation is essentially the same as that for credit risky bonds (Equation 13), but there is no maturity to the cash flows, so investors are essentially buying cash flows (dividends) into perpetuity (∞). In addition, we now have a new term in the discount rate, $\kappa_{t,s}^i$, which is the additional return that investors require for investing in equities, over and above what they require for investing in credit risky bonds $\left(l_{t,s} + \theta_{t,s} + \pi_{t,s} + \gamma_{t,s}^i\right)$.

The term $\kappa_{t,s}^i$ is essentially the equity premium relative to credit risky bonds. This is not the way the equity risk premium is usually expressed. We have expressed it this way for the moment because of the following reasons:

- If a company experiences financial difficulties because the company's debt holders have the senior claim on the company's cash flow, the equity holders will receive the residue, which could be zero, and
- in the event that a company's financial difficulties become so bad that the company is forced into bankruptcy, both bond and equity investors will lose. But depending on the quality of the company, investors in the corporate bond can usually expect to get some of their investment back. The equity investors, however, will normally lose all of their investment. Both debt and equity investors are exposed to risk, but the potential loss is greater for the equity investor.

These are the reasons why investors will require a risk premium, $\kappa_{t,s}^i$, over and above the one that they would require on the corporation's debt, $\gamma_{t,s}^i$. Equation 15 shows that both corporate bond and equity holders face what we might define as corporate risk. It is the combination of this risk that is usually referred to as the equity risk premium. Because the risk associated with equities is normally expressed relative to default-free debt of the same currency, we can rewrite Equation 15 as follows:

$$P_t^i = \sum_{s=1}^{\infty} \frac{E_t\left[\widetilde{CF}_{t+s}^i\right]}{\left(1 + l_{t,s} + \theta_{t,s} + \pi_{t,s} + \lambda_{t,s}^i\right)^s} \quad (16)$$

where the equity risk premium, $\lambda_{t,s}^i$, is equal to $\gamma_{t,s}^i + \kappa_{t,s}^i$. That is, it is the addition to return required by investors over and above the compensation for risk that they require for holding a default-free government bond of the same currency (technically, a very long-dated, plain-vanilla, coupon-paying, default-free bond).

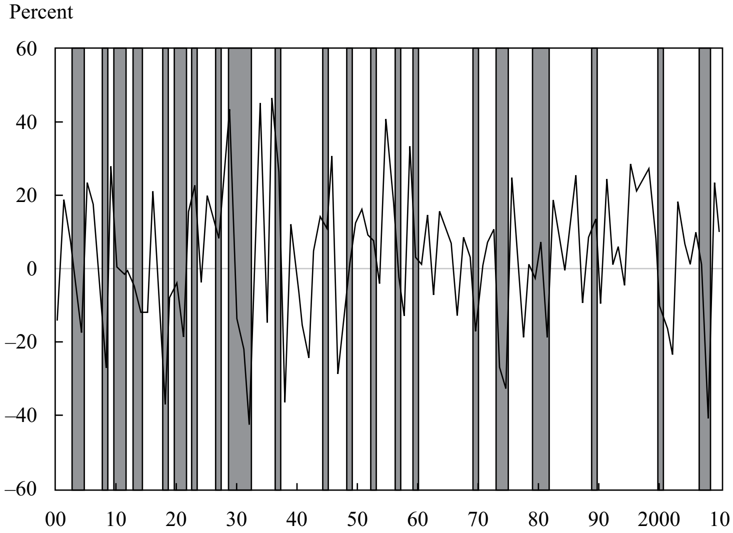
Equities and Bad Consumption Outcomes

Equity investors will demand an equity risk premium if the consumption hedging properties of equities are poor—that is, if equities tend not to pay off in bad times. Our arguments earlier indicate that the equity risk premium should be positive and therefore, implicitly, that equities are a bad hedge for bad consumption outcomes.

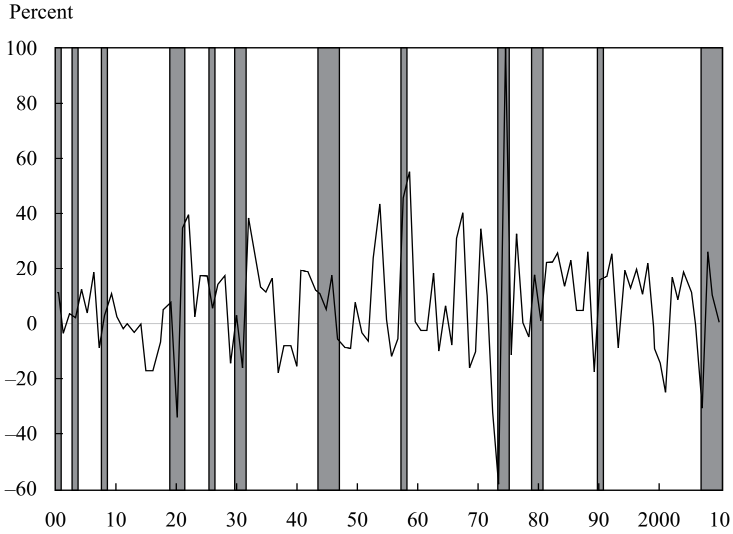
However, tying down the exact relationship between equity performance and consumption over time has proved to be very difficult. But we can get some idea of the relationship if we consider a very long history of the real returns produced by equities.

Exhibit 21: Annual Real Equity Returns, 1900–2010

United States



United Kingdom



Note: Shaded areas indicate recessions.

Sources: Based on data from Shiller (2000), the Bank of England, and the authors' calculations.

Exhibit 21 shows the annual real (inflation-adjusted) returns generated by both US (Panel A) and UK (Panel B) equities from 1900 to 2010. Generally speaking, sharp falls in equity prices are associated with recessions—bad times. For example, real UK equity prices more than halved as a recession hit the United Kingdom in 1972, and real US equity prices fell by more than 40% during the Great Depression. More recently, real UK and US equity prices fell by 30% and 40%, respectively, in 2009. Given this evidence, it is difficult to argue that equities are a good hedge for bad consumption outcomes. We would thus expect the equity risk premium to be positive, and given the scale of the declines in prices possible in bad times, we might expect it to be quite large.

Before we consider how large the equity premium should be, we will first focus on the cash flow that equities generate. It is the nature of this cash flow that leads investors to demand an equity risk premium in the first place.

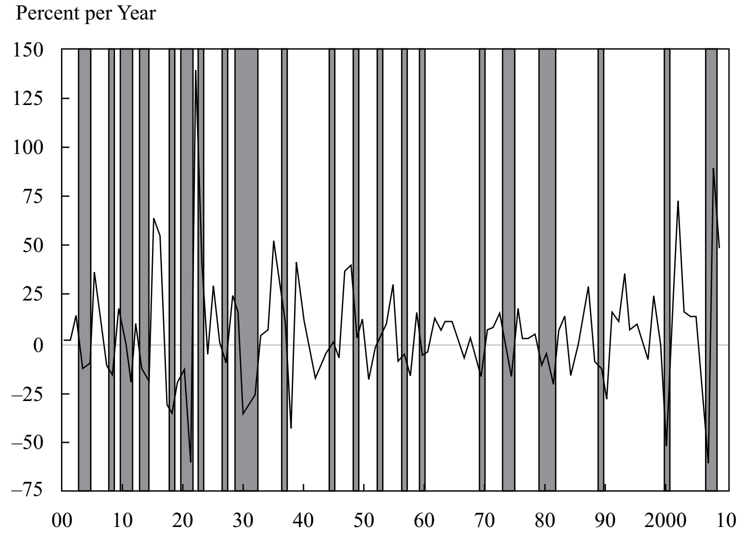
20 EARNINGS GROWTH AND THE ECONOMIC CYCLE

☐ explain how the phase of the business cycle affects short-term and long-term earnings growth expectations

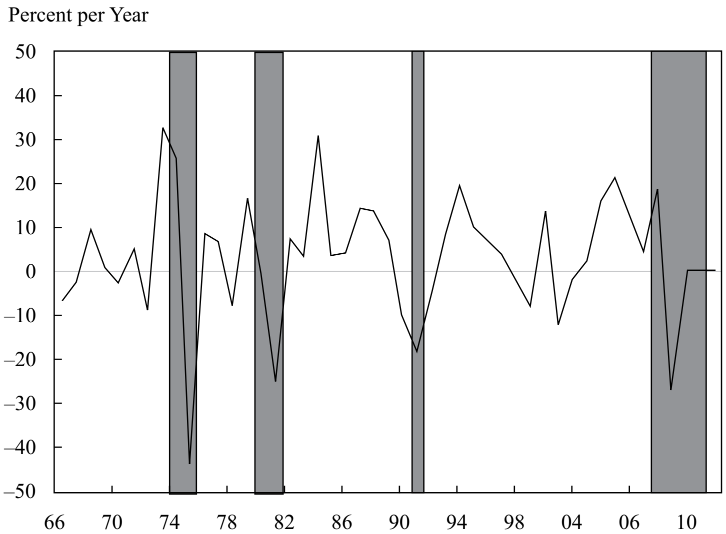
The uncertainty about—and time variation in—future dividends, as represented by the numerator in Equation 15, is a key feature of equity investment. Panels A and B of Exhibit 22 show a long history of US real earnings growth and a shorter history of UK real earnings growth, respectively. The exhibit shows that a sharp decline in real earnings nearly always coincides with a recession, which is to be expected; recessions are associated with declines in employment, incomes, output, and, subsequently, profitability. US real earnings fell dramatically during the Great Depression and by nearly 60% in 2009. Conversely, sharp increases in profit growth occur at the end of a period of recession and in some cases while recession conditions still persist. Thus, corporate profitability can lead an economy out of recession as well as into it: A negative demand shock can cause demand and corporate profits to shrink. In response, companies lay off workers, reducing their cost base and thereby adding to the recessionary backdrop. When an upturn in demand occurs, perhaps in response to monetary policy stimulus, demand growth on a lower cost base can lead to a sharp increase in corporate profits, which then leads companies to invest and hire more staff, and so on. Some analysts thus consider corporate profitability to be an important leading indicator of the business cycle and believe it provides useful information about future growth.

Exhibit 22: Real Equity Earnings Growth in the United States and the United Kingdom

United States



United Kingdom



Note: Shaded areas indicate recessions.

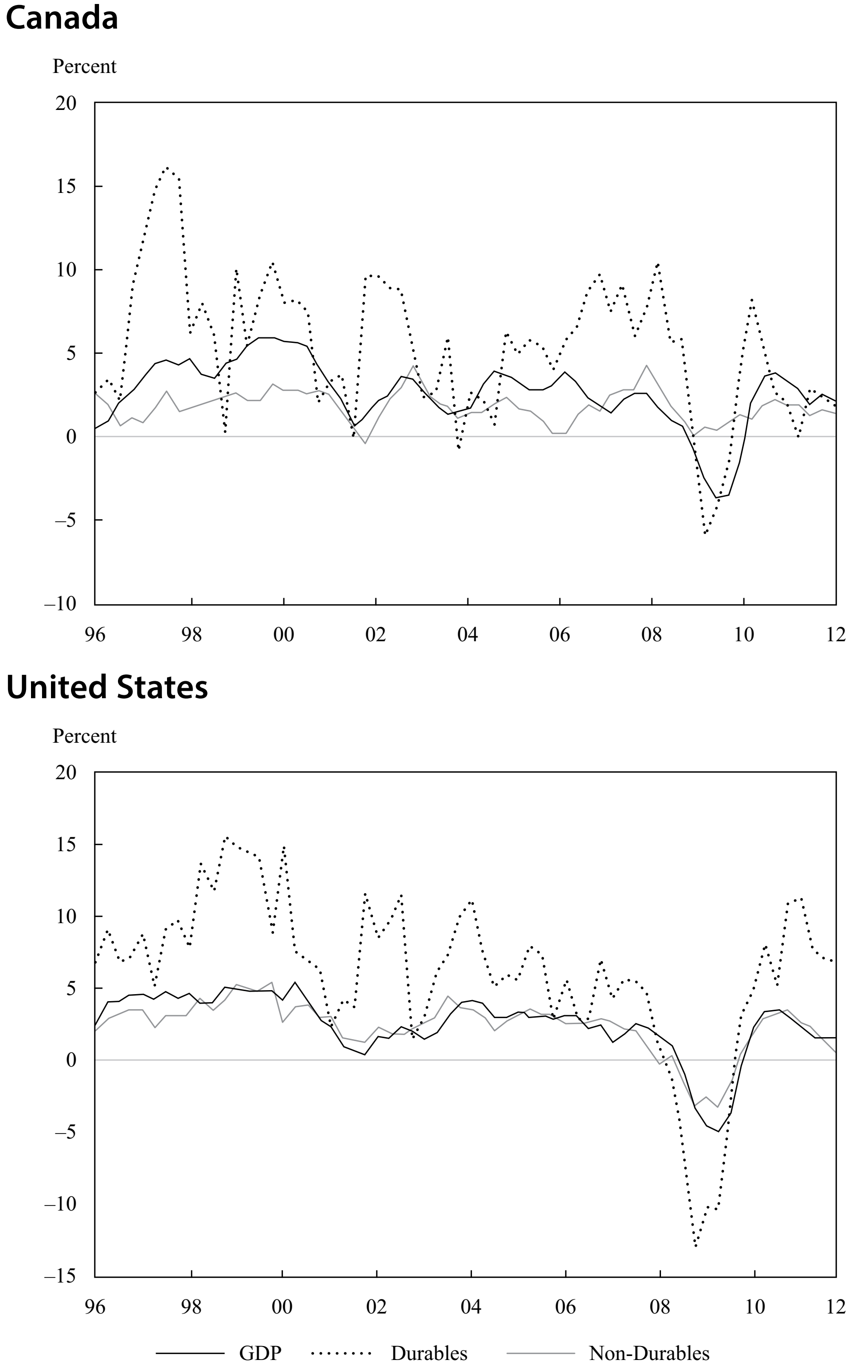
Sources: Based on data from Shiller (2000) and the Bank of England.

Equity analysts spend the majority of their time focusing on the numerator in Equation 15—that is, forming views about expected earnings and, therefore, about dividends and free cash flow. Given the close relationship between aggregate earnings, or profits, and the business cycle shown in Exhibit 22, an understanding of the business cycle is crucial for earnings projections, particularly in the short term. However, the business cycle will not affect the corporate profits of every company in the same way. The type of product sold or service provided by the company will have an impact on earnings and consequently on equity performance over the business cycle.

Some companies make products or provide services that are relatively insensitive to general economic conditions. Toothpaste might fall into this category. Because the cost of toothpaste usually only represents a small proportion of the overall household budget, people will generally still want to keep their teeth clean even if the economy is in recession, and because they are unlikely to want to clean their teeth more often simply because the economy is booming, the demand for toothpaste will remain fairly stable over the business cycle. Companies and equity sectors that produce such products are referred to as non-cyclical or defensive investments. By contrast, some companies, such as airlines, will produce goods or provide services that are extremely sensitive to the business cycle. In difficult economic conditions, consumers are much more likely to postpone or cancel their vacations or to vacation at home than to reduce their consumption of toothpaste, and businesses are likely to cut back on airline travel. Generally speaking, an annual family vacation will constitute a large proportion of the household budget, and most people do not need a vacation in the same way that they need toothpaste or soap. Businesses may rely on alternatives to expensive travel for meetings, such as video conferencing. By contrast, in good times when real incomes are rising, people are more likely to take more vacations or more expensive ones, and the increase in business activity may necessitate more meetings in new, often distant markets. Economists and investment strategists may view a rise in the earnings of cyclical companies after a period of decline as an indicator of a likely improvement in wider economic growth in the future.

Exhibit 23 shows the annual growth rates (year over year) of real GDP and of the consumption of both durable and non-durable goods for Canada (Panel A) and the United States (Panel B). Both panels of the exhibit show how sensitive durable goods consumption is to the economic cycle. We can expect then that the profits of companies that produce durable as opposed to non-durable goods to be commensurately more volatile too.

Exhibit 23: Year-over-Year Growth Rate of GDP and the Consumption of Durable and Non-Durable Goods, 1996–2012

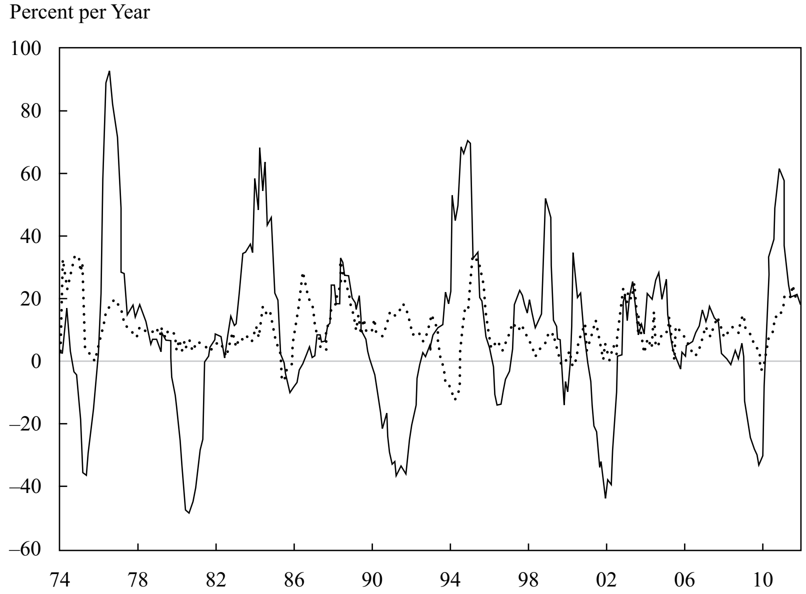


Source: Based on data from Thomson Reuters.

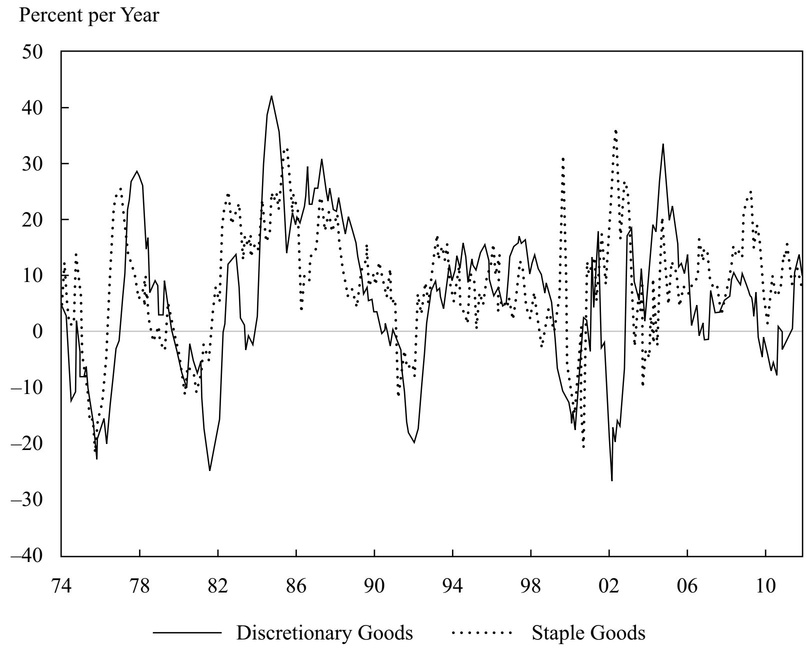
Stock market participants often classify stocks as being cyclical and non-cyclical. Exhibit 24 shows the real earnings growth of the non-cyclical and cyclical goods sectors of both the United States (Panel A) and the United Kingdom (Panel B). The cyclical sectors in this case are represented by companies that produce discretionary consumer goods, whereas the non-cyclical index is represented by an index that includes companies that produce staple (or less discretionary) consumer goods. Panel A shows the clearest evidence of the greater sensitivity to business conditions of the cyclical sector. Real earnings growth rises and falls dramatically over the business cycle. By contrast, although the real earnings of non-cyclical companies vary across the business cycle, the peaks and troughs are less extreme. The time variation in UK real earnings over the business cycle is also evident from the exhibit. However, the difference between the real earnings growth of the UK's cyclical and non-cyclical sectors is less clear, although the cyclical sector tended to experience more significant troughs in real earnings growth over this period.

Exhibit 24: The Real Earnings Growth of Discretionary and Staple Consumer Goods Companies, 1974–2012

United States



United Kingdom



Source: Based on data from Thomson Reuters.

There are, of course, other factors that determine the earnings growth of an equity or equity sector: the financial structure of the company, the quality and experience of its management, and the ease with which new entrants can establish themselves to compete away any abnormal profits. However, the relationship between the business cycle and the nature of the type of good or service sold will remain important. Indeed, in a booming economy, even bad managers of companies with poor financial structures can generate or appear to generate profits—for example, WorldCom and Enron. But tougher, recessionary conditions often expose weak companies as demand turns down and financing becomes harder to access.

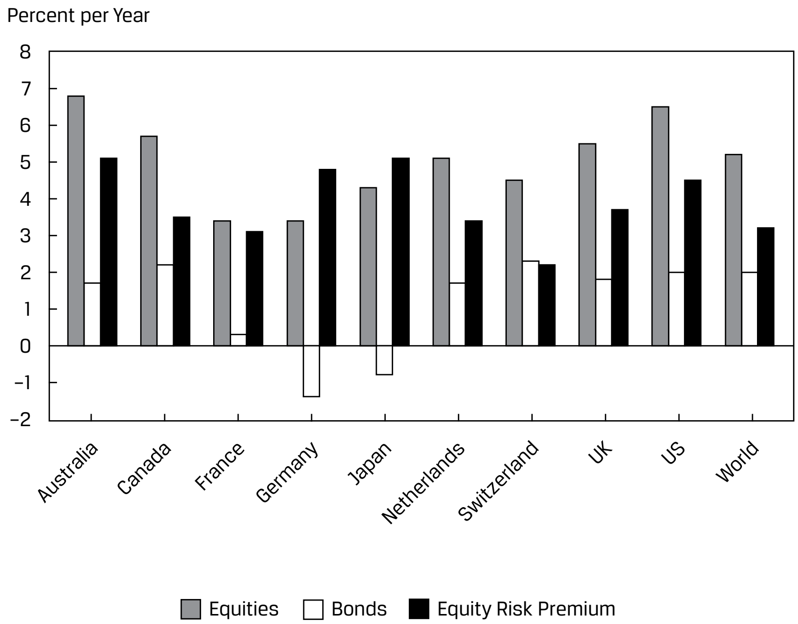
21 HOW BIG IS THE EQUITY RISK PREMIUM?

☐ explain the relationship between the consumption hedging properties of equity and the equity risk premium

The real earnings of companies are clearly affected by the underlying economy. This relationship is positive in that when the economy turns down, so (normally) do corporate profits. But it is in these bad times that investors need their investments to offset these worsening earnings. Because of the pro-cyclicality of economies and corporate profits (in aggregate), equities are not a good hedge against bad consumption outcomes, which, in turn, means that investors will require a risk premium. But how big should this premium $\left(\lambda_{t,s}^{i}\right)$ be?

It is impossible to quantify the equity risk premium *ex ante*. But we can at least look at its *ex post* value using very long runs of data. Exhibit 25 shows the real annual return on equities and government bonds over the period of 1900–2017 for a range of developed-economy equity markets. Over this very long period, equities in each country have outperformed government bonds. The bars representing the *ex post* equity risk premium range from 2.2% per year in Switzerland to 5.1% per year in Australia. US equities, which constitute the world's largest equity market, have outperformed US Treasuries by 4.5% per year on average over the 117 years under measurement, whereas the global equity market has produced an equity risk premium of 3.2% per year (premium versus bonds). Of course, there is no guarantee that a premium earned in the past will be earned in the future, but this long span of data shows that the *ex ante* equity risk premium for developed-economy equity markets could be somewhere between 3% and 5% per year.

Exhibit 25: *Ex-Post* Real Returns on Equities, Bonds, and Equity Risk Premiums



Source: Based on data from Chapter 1 of Dimson, Marsh, Staunton (2018).

VALUATION MULTIPLES

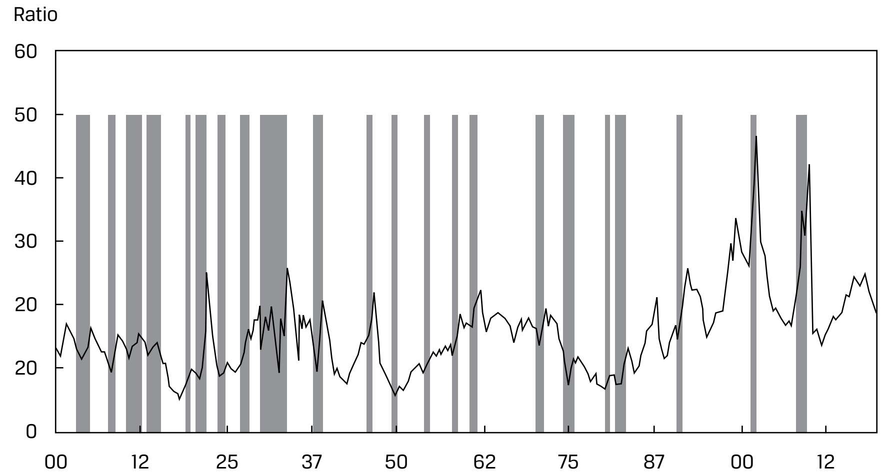
22

☐ describe cyclical effects on valuation multiples

Analysis of company earnings prospects is usually the central focus of equity analysts and strategists. To help compare equities within and among sectors, they will generally monitor valuation multiples, such as the price-to-earnings ratio (P/E) or the price-to-book ratio (P/B). P/E is calculated as the ratio of the current share price to the earnings per share (EPS) generated by the company. This ratio tells investors the price they are paying for the shares as a multiple of the company's earnings per share. Investors use this ratio to compare the valuations attributed to individual equities, sectors, and markets. For instance, if a stock is trading with a low P/E relative to the rest of the market, it implies that investors are not willing to pay a high price for a dollar's worth of the company's earnings. The reason may be that the market believes the prospect of strong earnings growth in the future is low. Alternatively, a share trading with a very high P/E relative to the rest of the market indicates that investors are willing to pay a higher price for each dollar's worth of the company's earnings. They may be willing to do so because they expect this company's earnings to grow rapidly in the future. When the EPS used to estimate the ratio refers to last year's earnings, the P/E is referred to as being a historical or a trailing P/E. However, when the EPS is based on an estimate of future earnings, it is referred to as the leading or forward P/E. If a company's EPS is expected to grow, then its historical P/E will be greater than its forward P/E. However, what constitutes a high or low P/E very much depends on the market, sector, or company in question and, in particular, on the economic backdrop. US P/Es between 1900 and 2018 are shown in Exhibit 26.

Another popular valuation multiple is the price-to-book ratio, which measures the ratio of the company's share price to its net assets or its assets minus liabilities attributed to each share. The P/B tells investors the extent to which the value of their shares is "covered" by the company's net assets. Some of these assets, such as office buildings, are tangible, whereas others, such as patents and copyrights, are intangible. Furthermore, some of the assets are actually on the balance sheet and hence part of book value, whereas others are not. It also indicates the strength of investors' expectations about the company's ability to generate a high return on its net assets, adjusted for risk. The higher the ratio, the greater the expectations for growth but the lower the safety margin if things do not turn out as expected. Again, what constitutes a high or low P/B is determined by the market, sector, and stock in question.

Exhibit 26: US P/Es, 1900–2018



Note: Shaded areas indicate recessions.

Sources: Based on data from Shiller (2000) and www.econ.yale.edu/~shiller/.

One of the problems for equity strategists is to ascertain whether the P/E (or P/B) is high or low. The average trailing US P/E between 1900 and 1990 was 13.5, indicating that investors were willing to pay $13.5 for a dollar's worth of the previous year's earnings; by the late 1990s and early 2000s, they were willing to pay $45. The expansion in the US equity market's P/E during the 1990s was a global phenomenon. Equity strategists justified the rise in the price of earnings with many ad hoc explanations—for example, the end of the cold war, better macro-policy that would ensure that major recessions were a thing of the past, and the internet revolution, to name just a few. However, with regard to the basic pricing relationship, shown in Equation 15, the high P/E could be the result of a number of factors, including

a. an increase in expectation of future real earnings growth $\left(E_t\left[\widetilde{CF}_{t+s}^{i}\right]\right)$;

b. falling real interest rates ($l_{t,s}$), possibly associated with falling volatility in real GDP growth;

c. a decrease in inflation expectations ($\theta_{t,s}$);

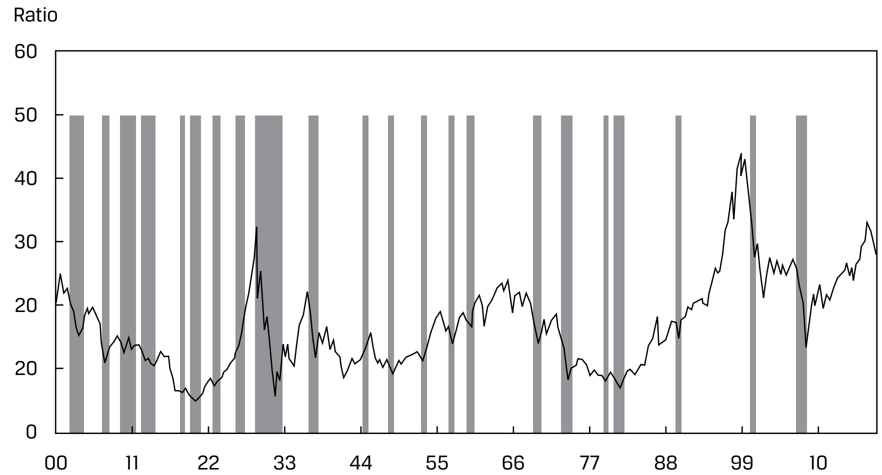
d. a decline in uncertainty about future inflation ($\pi_{t,s}$); or

e. a decrease in the equity risk premium $\left(\lambda_{t,s}^{i}\right)$.

Other things being equal, any one of these changes or all of them combined could justify higher equity prices (P) relative to current earnings (E) and thus higher equilibrium P/Es. There were some investors, however, who were not convinced that such high P/Es relative to historical levels were justifiable, particularly on the grounds of much higher future earnings growth. The US Federal Reserve Board chairman, Alan Greenspan, alarmed by the rise in P/Es, described the valuation of equity markets in 1996 as essentially the result of "irrational exuberance" on the part of equity investors.

Robert Shiller has proposed an alternative valuation multiple—the real cyclically adjusted P/E (CAPE). The CAPE is derived in the same way as the P/E, but the "P" represents the real (or inflation-adjusted) price of the equity market and the "E" is a 10-year moving average of the market's real (or inflation-adjusted) earnings. Deflating the real equity price by a moving average of real earnings irons out the short-term volatility in this indicator over time. Exhibit 27 shows this ratio for the United States from 1900 to 2018. The very high price that equity investors were willing to pay for equities in 1929 and 1999 and, to a lesser extent, in 1965 is still apparent. It is worth noting that the average real return on US equities in the 10 years following the peaks in the CAPE in 1929, 1965, and 1999 were −0.3% per year, −5.4% per year, and −4.1% per year, respectively. Conversely, the average real return over the 10-year period after the two lowest values of the CAPE in 1921 and 1980 were 12.3% per year and 7.3% per year, respectively.

Exhibit 27: Real US Cyclically Adjusted P/E (CAPE), 1900–2018



Note: Shaded areas indicate recessions.

Source: Based on data from Shiller (2000) and www.econ.yale.edu/~shiller/.

23 COMMERCIAL REAL ESTATE

☐ describe the economic factors affecting investment in commercial real estate

The basic pricing formula can be applied to asset classes besides bonds and equities. To demonstrate how the basic pricing framework presented in Equation 1 can be extended to other asset classes, we will consider commercial real estate.

Regular Cash Flow from Commercial Real Estate Investments

When investors invest in commercial real estate, the cash flow they hope to receive is derived from the rents paid by the tenants. These rents are normally collected net of ownership costs, such as those related to the upkeep of the building, according to a fixed schedule from the businesses that lease the property from the investors who act as landlords. Although practices vary from country to country, the rental agreement will be reviewed regularly and may be reset. In some countries, rents are subjected to "upward only" restrictions, which means that existing tenants will not see their rents fall, only potentially rise. Rents may also be indexed so that they rise in line with a pre-specified index of (usually) consumer prices.

Thus to a large extent, the rental income can be viewed as being analogous to the coupon income derived from a bond. Because a well-diversified portfolio of commercial property could be expected to generate a stream of rental income for investors, they might view such a portfolio as being similar to a well-diversified portfolio of bonds. The credit quality of a commercial property portfolio will be determined by the credit quality of the underlying tenants, in much the same way that the credit quality of a bond portfolio will be determined by the credit ratings of the bond issuers of the constituent bonds. Generally speaking, the lower the credit quality of the tenants, the less likely they will be to pay their rent on time or at all.

The Equity Component of an Investment in Commercial Real Estate

Investors in commercial real estate will receive regular cash flows derived from the rents paid by tenants, but there is another important element to property investment that is less bond-like. When a bond matures, the investor generally receives the face value of the bond along with the final coupon. But when the lease on a property expires, the investors (acting as landlords) will take back possession of the property and will have to decide whether to re-rent it to another tenant, to sell it to another investor, or to redevelop it for a future sale. The determining factor is likely to be the value of the property at the time. Its value may have risen dramatically over time, or it might now be worth much less. The value of the property will arguably be determined by two key factors: the property's location and the state of the underlying economy. If, during the time of the lease, the area in which the property is situated has become more popular, then the property might be sold at a profit or it might be worth redeveloping the property. Similarly, if the lease expires when general economic activity is high and thus there is strong demand for property, then the sale or redevelopment option might be worth pursuing. But if, when the lease expires, the location is deemed to be less desirable or the economy is weak, then redevelopment may not be an option, future rents may have to be lower on the property, and investors may come to the view that the property should be sold, even at a loss.

The potential for profit or loss and the uncertainty related to this profit from redevelopment add an equity-like dimension to investment in commercial real estate. In other words, this potential and uncertainty add either a positive increment to cash flow or a negative one. To this extent, some investors like to think about the cash flow derived from a commercial real estate portfolio as being part bond, part equity.

Illiquidity and Investment in Commercial Real Estate

There is a third aspect to investing in commercial real estate that is also crucial: its illiquidity. Anyone who has sold a home knows that it usually takes a great deal of time and effort to put the property up for sale, to find a buyer, and (if a buyer can be found) to finalize the deal. For similar reasons, it can take months and sometimes years to exit from a commercial property investment, and the high transactions costs often discourage investors further from liquidating holdings. By contrast, it is relatively easy in normal market conditions to transform a holding in developed-economy government bonds, investment-grade corporate debt, or publicly traded equities into cash. Generally speaking, most of the asset classes that we have considered so far in this reading are liquid relative to an investment in commercial property.

The Pricing Formula for Commercial Real Estate

Commercial real estate is a "special" asset class; it can be viewed as being part equity, part bond, and it is usually very illiquid. However, with some minor adaptions of the generic pricing formula in Equation 1, we can still capture all of the salient features of the price of commercial real estate, as follows:

$$P_t^i = \sum_{s=1}^{N} \frac{E_t\left(\widetilde{CF}_{t+s}^i\right)}{\left(1 + l_{t,s} + \theta_{t,s} + \pi_{t,s} + \gamma_{t,s}^i + \kappa_{t,s}^i + \phi_{t,s}^i\right)^s}. \tag{17}$$

The pricing formula shown in Equation 17 acknowledges that the expected cash flow from an investment in commercial real estate, $E_t\left(\widetilde{CF}_{t+s}^i\right)$, will be uncertain because tenants may default on the rental agreement. The quality of this rental income will depend on the quality of the tenants, just as the reliability or quality of the coupons from a corporate bond will be dependent on the credit standing of the corporate bond issuer. Furthermore, the property's value in the future cannot be known with certainty.

But what should the discount rate look like? To understand the construction of the discount rate in Equation 17, consider the following tenants and associated rental/leasing agreements:

1. a developed-economy government tenant that agrees to pay rental income that is indexed to inflation $(1 + l_{t,s})$,
2. a developed-economy government tenant that agrees to pay fixed nominal rental income $(1 + l_{t,s} + \theta_{t,s} + \pi_{t,s})$, and
3. a corporate tenant that agrees to pay a fixed nominal rental income $(1 + l_{t,s} + \theta_{t,s} + \pi_{t,s} + \gamma^i_{t,s})$.

In each case, the expressions in parentheses represent the composition of the discount rate that would be applied to the cash flows of bonds issued by these entities: (1) is analogous to the purchase of a real default-free government bond, (2) is analogous to the purchase of a nominal default-free government bond, and (3) is analogous to the purchase of a credit risky nominal bond. In each case, though, we need to add a risk premium to take into account the uncertainty relating to the value of the property at the end of the lease. This premium is analogous to the equity risk premium, $\kappa^i_{t,s}$.

Finally, we have to take into account the illiquidity of a commercial property investment. Because investors cannot easily convert their property investments into cash, there exists the possibility that they will not be able to liquidate their investment in bad economic times. In other words, other things being equal, illiquidity acts to reduce an asset class's usefulness as a hedge against bad consumption outcomes. Because of this, investors will demand a liquidity risk premium, which we have expressed as $\phi^i_{t,s}$ in Equation 17.

The discount rates that investors would apply to an investment in commercial property in each of the three instances previously listed are, therefore,

1. $1 + l_{t,s} + \kappa^i_{t,s} + \phi^i_{t,s}$,
2. $1 + l_{t,s} + \theta_{t,s} + \pi^i_{t,s} + \kappa^i_{t,s} + \phi^i_{t,s}$, and
3. $1 + l_{t,s} + \theta_{t,s} + \pi^i_{t,s} + \gamma^i_{t,s} + \kappa^i_{t,s} + \phi^i_{t,s}$.

The relative sizes of the components listed will vary depending on the length of the lease, the quality of the tenant, and the location of the property.

EXAMPLE 17

A Real Estate Investment Decision

1. An analyst estimates that the real risk-free rate is 1.25%, average inflation over the next year will be 2.5%, and the premium required by investors for inflation uncertainty is 0.50%. He also observes that the yield on a 10-year senior unsecured bond issued by Supermarket plc is 5.75%. From these figures, he deduces that the credit spread on Supermarket plc's 10-year debt is 1.50%.

   The same analyst is asked to review for a client (an investor) the opportunity to buy a site currently occupied by Supermarket plc. Once the investor purchases the property, Supermarket plc will lease it back and pay $500,000 annual rent in arrears to the investor. "Rent in arrears" in this case means that the first annual rental payment is due in 12 months, covering the first year's tenancy, and the second is due in 24 months, and so on. Like the Supermarket bond, the lease on the property has 10 full years to expire. At the end of this period, the property and land will revert to the investor, and the analyst estimates that the resale value of the property after 10 years will be $10 million, net of all transactions costs.

The investor tells the analyst that it normally expects to receive a risk premium of 0.50% on any cash flow from a commercial property investment to compensate it for the uncertainty of the final value of the property and the uncertainty relating to the receipt of rental income, plus a liquidity premium of 1.0% on these cash flows. The investor's required return on the property is thus 7.25% (= 5.75% + 0.5% + 1.0%). If the purchase price of this piece of commercial property is $8.2 million, should the analyst recommend the purchase to the client?

Discount Rate: 7.25%

Payment Due (years)	Cash Flow	Present Value
1	$500,000	$466,200
2	500,000	434,686
3	500,000	405,301
4	500,000	377,903
5	500,000	352,357
6	500,000	328,538
7	500,000	306,330
8	500,000	285,622
9	500,000	266,314
10	10,500,000	5,214,543
Implied property value		$8,437,796

The cash flows in the table, along with their associated present values, demonstrate that at a discount rate of 7.25%, the property would be priced at $8,437,796. Any asking price above this value would imply a return of less than the investor's hurdle rate of 7.25%, whereas any price below this price implies a return above this hurdle rate. On the basis of this information, the analyst should recommend that the client go ahead with the investment.

24 COMMERCIAL REAL ESTATE AND THE BUSINESS CYCLE

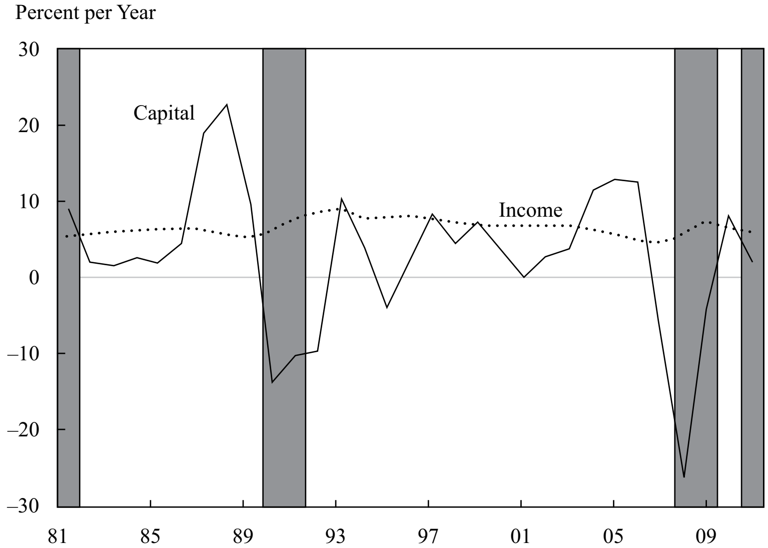
☐ describe the economic factors affecting investment in commercial real estate

The nature of the cash flows from commercial property and the complex structure of the discount rate will all be influenced by the evolution of the underlying economy. Panel A of Exhibit 28 shows the annual growth rate of UK commercial property income over a 30-year period. It is remarkably stable over this period, averaging 6.5% per year; in other words, UK commercial property rental income has grown by approximately

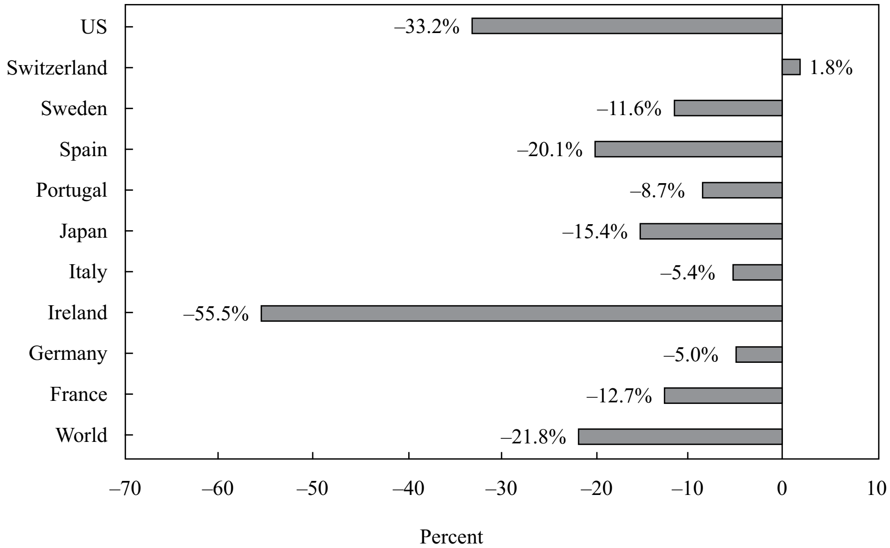
6.5% annually in nominal terms (or 2.5% in real terms) over the 30-year period. The stability of this income stream, over a number of business cycles, suggests that investors might calculate its present value using a very low discount rate. But as well as showing the annual change in rent from a portfolio of UK commercial property, Panel A also shows the annual percentage change in the capital value of the United Kingdom's commercial property market. Whereas rental income appears to have been relatively stable (in nominal terms) and almost immune to the business cycle, commercial property capital values are much more sensitive to the economic cycle. Between 1990 and 1992, as the UK economy experienced a deep recession, UK commercial property prices fell by a cumulative 30%. Over the course of the United Kingdom's 1990s recession, the capital value of the UK property market fell by 26%. In Panel B of the exhibit, we present the capital value changes in a number of markets around the world between 2008 and 2009. It is clear that the global recession had a significant impact on commercial property prices. For example, in Ireland, one of the developed economies arguably worst hit by the crisis, commercial property prices fell by 55.5%.

Exhibit 28: Commercial Property

UK Commercial Property Returns, 1981–2011



International Comparison of Commercial Property Value Changes, 2008–2009



Note: Shaded areas in Panel A indicate recessions.

Source: Based on data from Investment Property Databank (www.ipd.com).

Taken together, the two panels in Exhibit 28 show that even though nominal rental income might be relatively stable, the capital values of commercial property are highly sensitive to the economic environment. A recession will generally cause these values to fall, whereas more robust economic conditions will tend to cause commercial property prices to rise, often dramatically. For example, the recovering and then strong global economy between 2003 and 2006 caused world commercial property prices to rise by nearly 20%. Over the same period, they rose by 41% in the United Kingdom and by a staggering 51% in Ireland.

The pro-cyclical nature of commercial property prices means that investors will generally demand a relatively high risk premium in return for investing in this asset class. The reason is that commercial property does not appear to be a very good hedge against bad economic outcomes. In fact, the sharp declines in capital values in recessionary periods resemble the sort of declines that investors in equity experience, although these occurrences are more frequent with equity investment. Thus, the sort of risk premium that investors will demand from their commercial property investments arguably will be closer to that demanded on equities than on default-free government bonds.

Finally, although it is difficult to derive a value for the property risk premium, it is likely to vary over time with economic conditions and to be relatively highly and positively correlated with the risk premiums on corporate bonds and equities.

EXAMPLE 18

Valuation and the Business Cycle

1. Describe how real estate valuation is distinguished from valuation of public equities.

Solution

Real estate does not trade in public markets (the exception being REITs). Compared with the valuation of public equities, the valuation of real estate should reflect a discount for relative lack of liquidity.

SUMMARY

In this reading, we have sought to explain the fundamental connection between the prices of financial assets and the underlying economy. The connection should be strong because ultimately all financial assets represent a claim on the real economy. Because all financial assets offer a means of deferring consumption, to make the connection tangible we have explored the relationship between these asset prices and the consumption and saving decisions of economic agents.

- At any point in time, the market value of any financial security is simply the sum of discounted values of the cash flows that the security is expected to produce. The timing and magnitude of these expected cash flows will thus be an integral part of the security's market value, as will the discount rate applied to these expected cash flows, which is the sum of a real default-free interest rate, expected inflation, and possibly several risk premiums. Each of these elements will be influenced by the business cycle. It is through these components that the real economy exerts its influence on the market value of financial instruments.
- The average level of real short-term interest rates is positively related to the trend rate of growth of the underlying economy and also to the volatility of economic growth in the economy. Other things being equal, these relationships mean that we should expect to find that the average level of real short-term interest rates is higher in an economy with high and volatile growth and lower in an economy with lower, more stable growth.
- On average, over time, according to the Taylor rule, a central bank's policy rate should comprise the sum of an economy's trend growth plus inflation expectations, which might, in turn, be anchored to an explicit inflation target. This policy rate level is referred to as the neutral rate. Other things being equal, when inflation is above (below) the targeted level, the policy rate should be above (below) the neutral rate, and when the output gap is positive (negative), the policy rate should also be above (below) the neutral rate. The policy rate can thus vary over time with inflation expectations and the economy's output gap.
- Short-term nominal rates will be closely related to a central bank's policy rate of interest and will comprise the real interest rate that is required to balance the requirements of savers and investors plus investors' expectations

of inflation over the relevant borrowing or lending period. Short-term nominal interest rates will be positively related to short-term real interest rates and to inflation expectations.

- If bond investors were risk neutral, then the term structure of interest rates would be determined by short-term interest rate expectations. But bond investors are risk averse, which means that they will normally demand a risk premium for investing in even default-free government bonds. This risk premium will generally rise with the maturity of these bonds because longer-dated government bonds tend to be less negatively correlated with consumption and, therefore, represent a less useful consumption hedge for investors. Overall, the shape of the curve will be determined by a combination of short-term interest rates and inflation expectations as well as risk premiums. In turn, these factors will be influenced by the business cycle and policymakers.
- The yield differential between default-free conventional government bonds and index-linked equivalents will be driven by inflation expectations and a risk premium. The risk premium will be largely influenced by investors' uncertainty about future inflation.
- The difference between the yield on a corporate bond and that on a government bond with the same currency denomination and maturity is referred to as the measured credit spread. It is conceptually akin (but not equal) to the risk premium demanded by investors in compensation for the additional credit risk that they bear compared with that embodied in the default-free government bond. It tends to rise in times of economic weakness, as the probability of default rises, and tends to narrow in times of robust economic growth, when defaults are less common.
- The uncertainty about and time variation in future equity cash flows (dividends) is a distinct feature of equity investment, as opposed to corporate bond investment. This feature explains why we would expect the equity premium to be larger than the credit premium. In times of economic weakness or stress, the uncertainty about future dividends will tend to be higher, and we should thus expect the equity risk premium to rise in such an economic environment.
- Given the uncertain nature of the cash flows generated by equities, investors will demand an equity risk premium because the consumption hedging properties of equities are poor. In other words, equities tend not to pay off in bad times. Because in the event of company failure an equity holder will lose all of his or her investment whereas an investor in the company's bonds may recover a significant portion of his or her investment, it would be reasonable to assume that a risk-averse investor would demand a higher premium on an equity holding than on a corporate bond holding. The two premiums will tend to be positively correlated over time and will tend to be influenced by the business cycle in similar ways.
- The P/E tends to rise during periods of economic expansion and to fall during recessions. A "high" P/E could be the result of a number of factors, including the following: falling real interest rates, a decline in the equity risk premium, an increase in the expectation of future real earnings growth, an expectation of lower operating and/or financial risk, or a combination of all of these factors. All of these components will be influenced by the business cycle.

- The market value of an investment in commercial property can be derived in much the same way as the market value of an investment in equity. The cash flows come in the form of rent, which can be enhanced with additional redevelopment values as leases on properties expire. These cash flows are uncertain, and the uncertainty surrounding them will tend to rise when the economy turns down. We might thus expect the risk premium demanded on commercial property investments to rise in these times.
- The pro-cyclical nature of commercial property prices means that investors will generally demand a relatively high risk premium in return for investing in this asset class. The reason is that commercial property is not a very good hedge against bad economic outcomes. In addition, the illiquid nature of property investment means that investors may also demand a liquidity premium for investing in this asset class.

REFERENCES

Altug, Sumru and Pamela Labadie. 2008. *Asset Pricing for Dynamic Economies*. Cambridge, UK: Cambridge University Press.

Cochrane, John H. 2005. *Asset Pricing*. Princeton, NJ: Princeton University Press.

Dimson, Elroy, Paul Marsh, and Mike Staunton. 2018. *Credit Suisse Global Investment Returns Yearbook 2018*. Credit Suisse Research Institute.

Elton, Edward J., Martin J. Gruber, Stephen J. Brown, and William N. Goetzmann. 2014. *Modern Portfolio Theory and Investment Analysis*, 9th Ed. New York: John Wiley & Sons.

Shiller, Robert J. 2000. *Irrational Exuberance*. Princeton, NJ: Princeton University Press.

Taylor, John B. 1993. "Discretion versus Policy Rules in Practice." *Carnegie-Rochester Conference Series on Public Policy* 39, vol. : 195–214. 10.1016/0167-2231(93)90009-L

PRACTICE PROBLEMS

1. All else equal, which of the following would *mostlikely* explain the fall in price of a particular company's shares?
   - **A.** The expected inflation rate falls.
   - **B.** The company's future cash flows are expected to increase.
   - **C.** The yield to maturity on real default-free investments rises.

The following information relates to questions 2-8

Julie Carlisle is a financial planner at a large wealth management firm. One of her clients, Esteban Blake, just received a sizable inheritance. He invests a portion of the inheritance in an annuity that will immediately increase his income by a substantial amount. He enlists Carlisle's help to invest the remaining amount of the inheritance.

Blake informs Carlisle that he would like some short-term bonds in his portfolio. Carlisle proposes purchasing a one-year domestic government zero-coupon bond. It has a face value of $100 and is currently priced at $96.37. Carlisle estimates the one-year real risk-free rate at 1.15% and expects inflation over the next year to be 2.25%.

In an effort to provide Blake with some exposure to international markets, Carlisle proposes three countries to look for investment opportunities. Selected data on the three countries are presented in Exhibit 1.

Exhibit 1: Selected Macroeconomic Data

	Nominal GDP Growth	Inflation Rate	Volatility of Real GDP Growth	Yield Curve Shape	Trailing 12-Month Equity Index P/E
Country #1	6.5%	4.0%	Low	Flat	16.5
Country #2	5.0%	2.5%	High	Upward slope	17.3
Country #3	3.5%	2.0%	Low	Flat	18.2

In her analysis, Carlisle observes that the spread between the three-year default-free nominal bond and the default-free real zero-coupon bond in Country #3 is 2.0%.

Blake expresses concern that stocks may be currently overvalued in Country 3 given its 20-year historical equity index P/E of 16.0. Carlisle comments,

I think the equilibrium P/E in Country #3 has increased because of changes in market conditions.

Carlisle predicts that Country #3 will slip into a recession next quarter. She thinks it will be short-lived, lasting only 12 months or so, and considers the impact of such a recession on the performance of the country's stocks and bonds.

Exhibit 2: Three-Year Corporate Bonds from Country #3

Corporate Bond	Moody's Investors Service Rating	Spread*
Bond A	Aaa	1.4%
Bond B	Baa1	3.2%
Bond C	B3	5.3%

**Spread versus three-year sovereign bond*

2. Holding all else constant, the change in Blake's income will *most likely* result in:
   A. an increase in his marginal utility of consumption.
   B. an increase in his inter-temporal rate of substitution.
   C. a decrease in his required risk premium for investing in risky assets.

3. The implied premium for inflation uncertainty for the one-year government zero-coupon bond proposed by Carlisle is *closest* to:
   A. 0.23%.
   B. 0.37%.
   C. 1.10%.

4. Based on the data in Exhibit 1, current real short-term interest rates would *most likely* be highest in:
   A. Country #1.
   B. Country #2.
   C. Country #3.

5. The recent change in Country #3's break-even inflation rate suggests that the expected rate of inflation over the next three years is:
   A. less than 2.0%.
   B. equal to 2.0%.
   C. greater than 2.0%.

6. Which of the following changes in market conditions *best* supports Carlisle's comment regarding the equilibrium P/E for Country #3?
   A. An increase in the equity risk premium
   B. A decrease in uncertainty about future inflation
   C. A decrease in expectation of future real earnings growth

7. If Carlisle's prediction about the economy of Country #3 is realized, the yield curve in Country #3 will *most likely*:
   A. remain flat.
   B. become upward sloping.

C. become downward sloping.

8. Based on Exhibit 2, if Carlisle's prediction for Country #3 is realized, then over the next 12 months:

   A. Bond A would be expected to outperform Bond C.

   B. Bond B would be expected to outperform Bond A.

   C. Bond C would be expected to outperform Bond B.

---

9. The covariance between a risk-averse investor's inter-temporal rate of substitution and the expected future price of a risky asset is typically:

   A. negative.

   B. zero.

   C. positive.

10. The prices of one-period, real default-free government bonds are likely to be *most* sensitive to changes in:

    A. investors' inflation expectations.

    B. the expected volatility of economic growth.

    C. the covariance between investors' inter-temporal rates of substitution and the expected future prices of the bonds.

11. Default-free real interest rates tend to be relatively high in countries with high expected economic growth because investors:

    A. increase current borrowing.

    B. have high inter-temporal rates of substitution.

    C. have high uncertainty about levels of future consumption.

12. Positive output gaps are usually associated with:

    A. deflation.

    B. high unemployment.

    C. economic growth beyond sustainable capacity.

13. All else equal, an investor expects future inflation to increase, but the uncertainty of future inflation to fall. For such an investor the break-even inflation rate:

    A. is uncertain.

    B. is expected to fall.

    C. is expected to rise.

14. The difference between the yield on a zero-coupon, default-free nominal bond and the yield on a zero-coupon, default-free real bond of the same maturity

reflects:

A. investors' expectations about future inflation only.

B. a premium for the uncertainty of future inflation only.

C. both, investors' expectations about future inflation and a premium for the uncertainty of future inflation.

15. During a recession, the slope of the yield curve for default-free government bonds is *most likely* to:

A. flatten.

B. steepen.

C. become inverted.

16. One interpretation of an upward sloping yield curve is that the returns to short-dated bonds are:

A. uncorrelated with bad times.

B. more positively correlated with bad times than are returns to long-dated bonds.

C. more negatively correlated with bad times than are returns to long-dated bonds.

17. An analyst, who measures yield as a combination of interest rates and premiums, observes an upward-sloping, default-free government bond nominal yield curve. Which of the following statements is correct?

A. Interest rates must be expected to rise in the future.

B. Bond risk premiums must be expected to rise in the future.

C. Expectations relating to the future direction of interest rates are indeterminate.

18. A corporate bond has a remaining maturity of 1 year, has a face value of EUR100, and is currently priced at EUR90.90. The real risk-free rate is 3.25%. Inflation is expected to be 2.0% next year, and the premium required by investors for inflation uncertainty is 0.25%.

The implied credit risk premium embedded in the bond's price is *best* described as:

A. equal to (100/90.90) – 1 = 10%.

B. 10% reduced by the real risk-free rate and expected inflation.

C. 10% reduced by the real risk-free rate, expected inflation, and the premium for inflation uncertainty.

19. A decrease in the prices of AAA-rated corporate bonds during a recession would *mostlikely* be the result of:

A. expectations of higher inflation.

B. increases in credit risk premiums.

C. increases in short-term, default-free interest rates.

20. During an economic period when spreads between corporate and government bonds are narrowing, and spreads between higher- and lower-rated corporate bond categories are also narrowing, it can be expected that:

A. government bonds will outperform corporate bonds.

B. lower-rated corporate bonds will outperform higher-rated corporate bonds.

C. higher-rated corporate bonds will outperform lower-rated corporate bonds.

21. The sensitivity of a corporate bond's spread to changes in the business cycle is *most likely* to be:

A. uncorrelated with the level of cyclicality in the company's business.

B. positively correlated with the level of cyclicality in the company's business.

C. negatively correlated with the level of cyclicality of the company's business.

22. The category of bonds whose spreads can be expected to widen the *most* during an economic downturn are bonds from the:

A. cyclical sector with low credit ratings.

B. cyclical sector with high credit ratings.

C. non-cyclical sector with low credit ratings.

23. With regard to the credit risk of the sovereign debt issued by country governments, which of the following statements is correct? The credit risk premium on such debt is:

A. zero because governments can print money to settle their debt.

B. negligibly small because no country has defaulted on sovereign debt.

C. a non-zero and positive quantity that varies depending on a country's creditworthiness.

24. Risk-averse investors demanding a large equity risk premium are *most likely* expecting their future consumption outcomes and equity returns to be:

A. uncorrelated.

B. positively correlated.

C. negatively correlated.

25. Which of the following financial assets is likely to offer the *most* effective hedge against bad consumption outcomes?

A. Equities.

B. Short-dated, default-free government bonds.

C. Long-dated, default-free government bonds.

26. When assessing investment opportunities in equities, investors should:

A. assign higher equity risk premiums to non-cyclical companies, relative to cyclical companies.

B. forecast lower volatility in the growth rate of earnings for cyclical companies, relative to non-cyclical companies.

C. forecast higher growth rates in earnings for cyclical companies coming out of a recession, relative to non-cyclical companies.

27. Other things equal, equilibrium price-to-earnings ratios (P/Es) will *most likely* decrease if:

A. real interest rates decrease.

B. inflation is expected to increase.

C. there is less uncertainty about future inflation.

28. Which of the following statements relating to commercial real estate is correct?

A. Rental income from commercial real estate is generally unstable across business cycles.

B. Commercial real estate investments generally offer a good hedge against bad consumption outcomes.

C. The key difference in the discount rates applied to the cash flows of equity investments and commercial real estate investments relates to liquidity.

SOLUTIONS

1. C is correct. According to the fundamental pricing equation, the market value of an asset is affected by economic factors that influence the asset's expected future cash flows, default-free interest rates, expected inflation rates, or the asset's risk premium. From Equation 1, expected cash flows are in the numerator, while expected inflation and the real risk-free rate are in the denominator. Consequently, a rise in the real risk-free rate (the yield to maturity on a default-free instrument) will lead to a fall in the price of a risky asset, such as stock, by increasing the rate at which its cash flows are discounted.

2. C is correct. The additional annuity payment substantially increases Blake's income and wealth, which decreases his marginal utility of consumption. As a result, the average loss of marginal utility from any risk taking decreases as his wealth increases. Thus, he requires a lower risk premium and is willing to buy more risky assets.

3. B is correct. The pricing equation for a default-free nominal coupon-paying bond is

   $$P_t^i = \sum_{s=1}^{N} \frac{CF_{t+s}^i}{\left(1 + l_{t,s} + \theta_{t,s} + \pi_{t,s}\right)^s}.$$

   For a one-year bond, the pricing formula reduces to

   $$P_t = \frac{CF_{t+1}}{\left(1 + l_{t,1} + \theta_{t,1} + \pi_{t,1}\right)^1}.$$

   Thus, the implied premium for inflation uncertainty for the one-year government zero-coupon bond is calculated as

   $$\begin{aligned}\pi_{t,1} &= \frac{CF_{t+1}}{P_t} - \left(1 + l_{t,1} + \theta_{t,1}\right)\\ &= \frac{100}{96.37} - (1 + 0.0115 + 0.0225)\\ &= 1.0377 - 1.0340\\ &= 0.0037\text{, or } 0.37\%.\end{aligned}$$

4. B is correct. Real short-term interest rates are positively related to both real GDP growth and the volatility of real GDP growth. Country 1 and Country 2 have the highest real GDP growth, as estimated by the difference between nominal GDP growth and average inflation (6.5% – 4.0% = 2.5% and 5.0% – 2.5% = 2.5%, respectively), while Country 3 has the lowest real GDP growth (3.5% – 2.0% = 1.5%). Looking at the volatility of real GDP growth, Country 2 has high real GDP growth volatility, whereas Country 1 and Country 3 have low real GDP growth volatility. Therefore, Country 2 would most likely have the highest real short-term interest rates.

5. A is correct. The difference, or spread, between the yields on the country's three-year default-free nominal bond and on the default-free real zero-coupon bond is 2.0%. This spread is known as the break-even rate of inflation (BEI), which is composed of the expected rate of inflation plus a risk premium for the uncertainty of future inflation. Because this risk premium component is most likely positive, because investors are unlikely to be very confident in their ability to predict inflation accurately, the expected rate of inflation component would be

less than 2.0%.

6. B is correct. Stock prices are a function of expected cash flows discounted by inflation expectations, the uncertainty of future inflation, and the equity risk premium, among other factors. Holding all else equal, a decline in the uncertainty of future inflation would result in lower discount rates and higher valuations. This result would support a higher equilibrium P/E, thus justifying Country 3's current trailing P/E being higher than its historical average.

7. B is correct. The yield curve in Country 3 is currently flat (Exhibit 1), and Carlisle predicts a recession. During a recession, short-term rates tend to be lower because central banks tend to lower their policy rate in these times. However, the impact of monetary policy on longer-term rates will not be as strong because the central bank will usually be expected to bring short-term rates back to normal as the recession recedes. Thus, the slope of the yield curve will likely become upward sloping during the recession.

8. A is correct. If Country 3 experiences a recession over the next 12 months, the credit spreads for corporate bonds would be expected to widen as investors sell the low-quality debt of issuers with high default risk and trade up to the higher-quality debt of issuers with low default risk. The issuers with a good credit rating (such as Aaa rated Bond A) tend to outperform those with lower ratings (such as B3 rated Bond C) as the spread between low- and higher-quality issuers widens. As a result, Bond A would be expected to outperform Bond C over the next 12 months.

9. A is correct. For risk-averse investors, when the expected future price of the investment is high (low), the marginal utility of future consumption relative to that of current consumption is low (high). Hence, the covariance of the inter-temporal rate of substitution with asset price is expected to be negative for risk-averse investors.

10. B is correct. Only changes in default-free real interest rates will affect the price of real, default-free bonds. The average level of default-free real interest rates is positively related to the volatility of economic growth in the economy; thus, changes in the expected volatility of economic growth would likely lead to changes in default-free real interest rates, which in turn would affect the prices of real, default-free government bonds.

11. A is correct. The average level of default-free real interest rates is positively related to the expected rate of growth of the underlying economy and also to the volatility of economic growth in the economy. During periods of high expected economic growth, investors are less worried about the future and their consumption abilities in the future; that is, their inter-temporal rate of substitution is low, so they borrow more today and save less. Other things being equal, this means that the average level of default-free real interest rates (the reciprocal of the rate of substitution, see Equation 4) should be higher in an economy with high growth and lower in an economy with lower, more stable growth.

12. C is correct. An economy operating with a positive output gap—that is, where the level of actual GDP exceeds potential GDP—is producing beyond its sustainable capacity. Positive output gaps are usually associated with high and/or rising inflation, while high levels of unemployment usually accompany negative output gaps.

13. A is correct. The break-even inflation rate is the difference between the yield on a zero-coupon, default-free nominal bond and on a zero-coupon, default-free real

bond of the same maturity. The rate incorporates changing expectations about inflation and changing perceptions about the uncertainty of the future inflation environment. Consequently, if inflation is expected to rise while the uncertainty about future inflation falls (in Equation 10, $\theta_{t,s}$ rises but $\pi_{t,s}$ falls), it is unclear in which direction break-even inflation rates will move.

14. C is correct. The difference between the yield on a zero-coupon, default-free nominal bond and the yield on a zero-coupon, default-free real bond of the same maturity is known as the break-even inflation rate. This break-even inflation rate will incorporate the inflation expectations of investors over the investment horizon of the two bonds, plus a risk premium to compensate investors for uncertainty about future inflation. Break-even inflation rates are not simply the market's best estimate of future inflation over the relevant investment horizon, because break-even inflation rates also include a risk premium to compensate investors for their uncertainty about future inflation.

15. B is correct. During a recession, short rates are often lower because central banks tend to lower their policy rate in these times because the output gap is likely to be negative. However, the impact of such monetary policy on longer-term rates will not be as strong, so long rates may not fall by as much as short rates. The central bank will usually be expected to bring short-term rates back to normal as the recession recedes, and the risk-free rates will increase as economic growth recovers. Thus, the slope of the yield curve will typically steepen during a recession.

16. C is correct. One interpretation of an upward-sloping yield curve is that returns to short-dated bonds are more negatively correlated with bad times than are returns to long-dated bonds. This interpretation is based on the notion that investors are willing to pay a premium and accept a lower return for short-dated bonds if they believe that long-dated bonds are not a good hedge against economic "bad times."

17. C is correct. An upward-sloping yield curve may be caused by a combination of expected rate increases and positive bond risk premiums. It may also be a combination of expectations that interest rates will be unchanged in the future coupled with positive bond risk premiums. Lastly, an upward-sloping yield curve may actually be a reflection of expected rate cuts that are more than offset by the existence of positive bond risk premiums. So, expectations relating to the future direction of interest rates are indeterminate.

18. C is correct. The implied credit risk premium embedded in the bond's price is the yield (10%) less the default-risk-free nominal interest rate, which includes a premium for inflation uncertainty. See Example 15. The credit risk premium can be calculated as 4.51% in this case:

$$\gamma_{t,s}^{i} = \frac{100}{90.90} - (1 + 0.0325 + 0.02 + 0.0025).$$

$$\gamma_{t,s}^{i} = 4.51\%.$$

19. B is correct. During recessions, the risk premium that investors demand on financial assets, particularly those that are not default-free, such as corporate bonds, may rise because investors in general may be less willing and able to take on heightened default risk during such periods. Specifically, the credit risk premium demanded by investors tends to rise in times of economic weakness, when the probability of a corporate default and bankruptcy is highest.

20. B is correct. When spreads are narrowing, investors seem to be less discerning between issues with weak versus strong credit, and the rate of improvement

will tend to be greater for those bonds issued by entities with a relatively weaker ability to pay. Thus, during times when corporate bond spreads are narrowing relative to government bonds and the spreads between higher- and lower-rated bond categories are also narrowing, corporate bonds will generally outperform government bonds and lower-rated corporate bonds will tend to outperform higher-rated corporate bonds.

21. B is correct. The sensitivity of a corporate bond's spread to changes in the business cycle and the level of cyclicality tend to be positively correlated. The greater the level of cyclicality, the greater the sensitivity of the bond's spread to changes in the business cycle.

22. A is correct. During an economic downturn, the spreads of corporate bonds can be expected to widen, because the risk premium that investors demand on risky financial assets will increase. When spreads widen, the spreads on bonds issued by corporations with a low credit rating and that are part of the cyclical sector will tend to widen the most.

23. C is correct. Credit premiums have been an important component of the expected return on bonds issued by countries (sovereign debt). The credit premium varies from country to country depending on how creditworthy investors consider it to be. The fact that countries have both printed money to pay back debt and defaulted on it gives rise to a non-zero credit risk premium.

24. B is correct. If investors demand high equity risk premiums, they are likely expecting their future consumption and equity returns to be positively correlated. The positive correlation indicates that equities will exhibit poor hedging properties, because equity returns will be high (i.e., pay off) during "good times" and will be low (i.e., not pay off) during "bad times." In other words, the covariance between risk-averse investors' inter-temporal rates of substitution and the expected future prices of equities is highly negative, resulting in a positive and large equity risk premium. This is the case because in good times, when equity returns are high, the marginal value of consumption is low. Similarly, in bad times, when equity returns are low, the marginal value of consumption is high. Holding all else constant, the larger the magnitude of the negative covariance term, the larger the risk premium.

25. B is correct. The relative certainty about the real payoff from short-dated, default-free government bonds and, therefore, the relative certainty about the amount of consumption that the investor will be able to undertake with the payoff indicate that an investment in such bonds would be a good hedge against bad consumption outcomes.

26. C is correct. During recessions, cyclical companies are likely to experience sharp declines in earnings, more so than non-cyclical companies. In contrast, while coming out of a recession, cyclical companies are likely to generate higher earnings growth relative to non-cyclical companies.

27. B is correct. Other things being equal, an increase in inflation expectations would result in lower equity prices relative to current earnings. This would result in lower equilibrium P/Es.

28. C is correct. To arrive at an appropriate discount rate to be used to discount the cash flows from a commercial real estate investment, a liquidity premium is added to the discount rate applicable to equity investments. The added liquidity premium provides additional compensation for the risk that the real estate investment may be very illiquid in bad economic times.

LEARNING MODULE

2

Analysis of Active Portfolio Management

by Roger G. Clarke, PhD, Harindra de Silva, PhD, CFA, and Steven Thorley, PhD, CFA.

Roger G. Clarke, PhD (USA). Harindra de Silva, PhD, CFA, is at Analytic Investors, Wells Fargo Asset Management (USA). Steven Thorley, PhD, CFA, is Emeritus Faculty at the Marriott School, BYU (USA).

LEARNING OUTCOMES

Mastery	*The candidate should be able to:*
☐	describe how value added by active management is measured
☐	calculate and interpret the information ratio (ex post and ex ante) and contrast it to the Sharpe ratio
☐	describe and interpret the fundamental law of active portfolio management, including its component terms—transfer coefficient, information coefficient, breadth, and active risk (aggressiveness)
☐	explain how the information ratio may be useful in investment manager selection and choosing the level of active portfolio risk
☐	compare active management strategies, including market timing and security selection, and evaluate strategy changes in terms of the fundamental law of active management
☐	describe the practical strengths and limitations of the fundamental law of active management

1 INTRODUCTION

The Markowitz (1952) framework of what was originally called modern portfolio theory (MPT) has now become the prominent paradigm for communicating and applying principles of risk and return in portfolio management. Much of the mathematics and terminology of mean–variance portfolio theory was subsequently combined with the notion of informational efficiency by Sharpe (1964) and other financial economists to develop equilibrium models, such as the traditional capital asset pricing model. Separately, the tools of MPT were applied by Treynor and Black (1973) to guide investors in their selection of securities when prices differ from their equilibrium values. The application of portfolio theory to active management was further developed by Grinold (1989) in "The Fundamental Law of Active Management" and by Black and Litterman (1992).

We summarize the principles of active portfolio management using the terminology and mathematics of the fundamental law introduced by Grinold (1989) and further developed by Clarke, de Silva, and Thorley (2002). Active management theory deals with how an investor should construct a portfolio given an assumed competitive advantage or skill in predicting returns. Thus, active management relies on the assumption that financial markets are not perfectly efficient. Although investors might ultimately care about total risk and return, when asset management is delegated to professional investors in institutional settings (e.g., pension funds) the appropriate perspective is risk and return relative to a benchmark portfolio. In addition to the principal–agent problem in delegated asset management, the availability of passively managed portfolios requires a focus on value added above and beyond the alternative of a low-cost index fund.

We assume an understanding of basic portfolio theory, including the mathematics of expected values, variances, and correlation coefficients, as well as some familiarity with the related disciplines of mean–variance optimization and multi-factor risk models. The following sections introduce the mathematics of value added through active portfolio management, including the concepts of active weights, relative returns, and performance attribution systems. The subsequent section compares the well-known Sharpe ratio for measuring the total risk-adjusted value added with the information ratio for measuring relative risk-adjusted value added. This section also makes a distinction between *ex ante*, or expected, risk and return versus *ex post*, or realized, risk and return and explains that the information ratio is the best criterion for evaluating active investors. We then introduce the fundamental law that describes how relative skill, breadth of application, active management aggressiveness, and the constraints in portfolio construction combine to affect value added. The remaining sections provide examples of active portfolio management strategies in both the equity and fixed-income markets, describe some of the practical limitations of the fundamental law, and provide a summary of the concepts and principles.

2 ACTIVE MANAGEMENT AND VALUE ADDED

☐ describe how value added by active management is measured

The objective of active management is to add value in the investment process by doing better than a benchmark portfolio. Value added is a relative performance comparison to investing in the benchmark portfolio, often called passive investing. If the investor outperforms the benchmark portfolio, value added is positive. If the investor underperforms the benchmark portfolio, value added is negative. In the latter case, the investor would have been better off during the measurement period by simply holding the benchmark portfolio, particularly net of fees and expenses. Examples of indexes that are used as benchmark portfolios include the MSCI All Country World Index and the Bloomberg Barclays Global Aggregate Bond Index, which represent the performance of global equities and global bonds, respectively.

Choice of Benchmark

A benchmark or passive portfolio should have a number of qualities to serve as a relevant comparison for active management:

- The benchmark is representative of the assets from which the investor will select.
- Positions in the benchmark portfolio can actually be replicated at low cost.
- Benchmark weights are verifiable *ex ante*, and return data are timely *ex post*.

An available security market index is often used as the benchmark portfolio. The most common market indexes weight the individual assets by their market capitalization. Capitalization weighting has played a prominent role in the development of capital market theory because such indexes are generally self-rebalancing and can be simultaneously held by many investors. Float-adjusted market capitalization-weighted indexes represent an incremental improvement over non-float-adjusted indexes by accounting for the percentage of a security or asset that is not privately held and thus available to the general investing public. One important consequence of using a float-adjusted capitalization-weighted market index as the benchmark is that when all relevant assets are included in the market, the value added from active management becomes a zero-sum game with respect to the market. Because the market portfolio represents the average performance across all investors that own securities before costs, active investors as a group cannot outperform the market (i.e., active management is a zero-sum game). For benchmarks that have a narrower definition than the total market, active management is not a zero-sum game because investors can select assets outside the benchmark.

The return on the benchmark portfolio, R_B, is based on the returns to the individual securities and the weights of each security in the portfolio:

$$R_B = \sum_{i=1}^{N} w_{B,i} R_i, \tag{1}$$

where R_i is the return on security i, $w_{B,i}$ is the benchmark weight of security i, and N is the number of securities. Similarly, the return on an actively managed portfolio, R_P, is a function of the weights of the securities, i, held in the portfolio, $w_{P,i}$, and the returns to the individual securities:

$$R_P = \sum_{i=1}^{N} w_{P,i} R_i. \tag{2}$$

The benchmark might include securities that are not part of the actively managed portfolio and thus would have a weight of zero by definition or simply be left out of the calculation in Equation 2. Similarly, an investor could include securities in the active portfolio that are not in the benchmark, and those would have a benchmark weight of zero in Equation 1. Please note that for simplicity, the same notation, N, is used in the summation in the expression for the managed portfolio return and the benchmark return, although fewer or more securities may be in the managed portfolio than in the benchmark.

Measuring Value Added

The value added or "active return" of an actively managed portfolio is typically calculated as the simple difference between the return on that portfolio and the return on the benchmark portfolio,

$$R_A = R_P - R_B,$$

and can thus be either positive or negative. A risk-adjusted calculation of value added, which we will refer to as the managed portfolio's alpha, incorporates some estimate of the managed portfolio's risk relative to the benchmark, often captured by the portfolio's beta, $\alpha_P = R_P - \beta_P R_B$. Unfortunately, the term *alpha* in practice is often used to refer to active return as well, which implicitly assumes that the beta of the managed portfolio relative to the benchmark is 1.

Equations 1 and 2 can be combined to illustrate the important principle that value added is ultimately driven by the differences in managed portfolio weights and benchmark weights: $\Delta w_i = w_{P,i} - w_{B,i}$. These values are called the active weights of the managed portfolio, and the symbol Δ (Greek letter delta) is used to indicate the difference from the benchmark weights. Combining Equations 1 and 2 and employing this definition for active weights yields the conceptually important result that value added is the sum product of the active weights and asset returns:

$$R_A = \sum_{i=1}^{N} \Delta w_i R_i.$$

Given that the sum of the active weights is zero, we can also write the value added as the sum product of active weights and active security returns:

$$R_A = \sum_{i=1}^{N} \Delta w_i R_{Ai}, \quad (3)$$

where $R_{Ai} = R_i - R_B$. Equation 3 indicates that positive value added is generated when securities that have returns greater than the benchmark are overweighted and securities that have returns less than the benchmark are underweighted.

Whereas many applications of value added focus on individual securities as the assets, we first illustrate the concept with a simple numerical example of a composite portfolio that has just two assets—a stock portfolio and a bond portfolio. Suppose the benchmark is a 60/40 weighted composite portfolio of stocks and bonds. The investor believes that over the next year stocks will outperform bonds, so the investor holds a portfolio that is weighted 70% stocks and 30% bonds. The managed portfolio is said to be *overweight* stocks by 10 percentage points and *underweight* bonds by 10 percentage points—in other words, an active weight of −10 percentage points on bonds. Assume that *ex post* (i.e., "after the fact"), the return on the stock market turned out to be 14.0% and the return on the bond market turned out to be just 2.0%. In this case, the return on the managed portfolio is 0.70(14.0) + 0.30(2.0) = 10.4% and the return on the benchmark is 0.60(14.0) + 0.40(2.0) = 9.2%.

From these final numbers, one could directly calculate the value added as 10.4 − 9.2 = 1.2%. But using Equation 3, a more informative calculation of value added showing the contributions from each segment is R_A = 0.10(14.0 − 9.2) − 0.10(2.0 − 9.2) = 0.5 + 0.7 = 1.2%. This breakout suggests that a 0.5% return relative to the benchmark was generated by being overweight stocks, and a 0.7% return was generated simultaneously by being underweight bonds—for a total of 1.2%. Of course, the actual returns might have been different—with the stock market return being lower than the bond market return, resulting in negative value added in the managed portfolio. For example, if the stock market had a return of −14.0% instead of +14.0%, the portfolio return and benchmark return would have been −9.2% and −7.6%, respectively. Then the value added from this single overweight/underweight decision would have been R_A = 0.10(−14.0) − 0.10(2.0) = −1.4% − 0.2% = −1.6%.

EXAMPLE 1

Value Added and Country Equity Markets

Consider the MSCI EAFE Index as the benchmark for an actively managed portfolio that includes allocations to individual countries, as given in the following exhibit. The portfolio (both benchmark and managed) weights are for the beginning of 2018. The portfolio manager actively changes country allocations but does not engage in security selection.

Country	Benchmark Weight	Portfolio Weight	2018 Return
United Kingdom	17%	16%	−7.6%
Japan	25%	14%	−9.0%
France	11%	8%	−3.5%
Germany	9%	24%	−15.8%
Other Countries	38%	38%	−0.1%

Source: Data from MSCI.

1. Which countries have the largest overweight and largest underweight in the managed portfolio compared with the benchmark portfolio? What are the active weights for these two countries?

Solution:

Germany has the largest overweight at 24 − 9 = +15%, and Japan has the largest underweight at 14 − 25 = −11%.

2. Using active weights and total returns, what was the value added of the managed portfolio over the benchmark portfolio in the calendar year 2018?

Solution:

The value added is −0.01(−7.6) − 0.11(−9.0) − 0.03(−3.5) + 0.15(−15.8) = −1.2%. Note that the "Other Countries" active weight is zero, so this asset does not contribute anything to the portfolio's active return. The value added can also be calculated using relative returns in Equation 3 with the same net result.

Decomposition of Value Added

In contrast to the previous simple example, performance attribution systems often attempt to decompose the value added into *multiple* sources. The most common decomposition is between value added due to asset allocation and value added due to security selection. Consider a composite portfolio of stocks and bonds where the asset allocation weights differ from a composite benchmark *and* each asset class is actively managed by selecting individual securities. The total value added is the difference between the actual portfolio return and the benchmark return:

$$R_A = \sum_{j=1}^{M} w_{P,j} R_{P,j} - \sum_{j=1}^{M} w_{B,j} R_{B,j}.$$

The first summation has both portfolio weights and the returns on actively managed portfolios, designated by the "*P*" subscript. The second summation has both benchmark weights and benchmark returns, designated by the "*B*" subscript. The subscript $j = 1$ to M counts the number of asset classes, leaving the notation subscript $i = 1$ to N for use elsewhere to count the securities within each asset class.

We can rewrite the total value added as the sum of the active asset allocation decisions and the weighted sum of the value added from security selection, $R_{A,j} = R_{P,j} - R_{B,j}$, within each asset class:

$$R_A = \sum_{j=1}^{M} \Delta w_j R_{B,j} + \sum_{j=1}^{M} w_{P,j} R_{A,j}, \tag{4}$$

although this formulation arbitrarily assigns an interactive effect to security selection. The performance attribution system in Equation 4 may be easier to conceptualize with just two asset classes, stocks and bonds (in other words, with $M = 2$). Using *stocks* and *bonds* as the subscripts, Equation 4 becomes:

$$R_A = \left(\Delta w_{stocks} R_{B,stocks} + \Delta w_{bonds} R_{B,bonds}\right) + \left(w_{P,stocks} R_{A,stocks} + w_{P,bonds} R_{A,bonds}\right).$$

The first (parenthetical) term is the value added from the asset allocation decision. The second term is the value added from security selection within the stock and bond portfolios. The active weights in the first term refer to differences from the policy portfolio. For example, the long-term policy portfolio might be 60/40 stocks versus bonds, and the investor deviates from this policy portfolio from year to year based on beliefs about the returns to each asset class.

To give a numerical example, consider the fund returns for the calendar year 2018 in the following table.

Fund	Fund Return (%)	Benchmark Return (%)	Value Added (%)
Fidelity Magellan	−5.6	−4.5	−1.1
PIMCO Total Return	−0.3	0.0	−0.3
Portfolio Return	−3.9	−2.7	−1.2

Specifically, the Fidelity Magellan mutual fund had a return of −5.6%, compared with a −4.5% return for its benchmark, the S&P 500 Index. In the same year, the PIMCO Total Return Fund had a return of −0.3%, compared with a 0.0% return for its benchmark, the Bloomberg Barclays US Aggregate Index. Consider an investor who invested in both actively managed funds, with 68% of the total portfolio in Fidelity and 32% in PIMCO. Assume that the investor's policy portfolio (strategic asset allocation) specifies weights of 60% for equities and 40% for bonds.

- As shown in the table, Fidelity Magellan added value of $R_A = R_P - R_B =$ −5.6% − (−4.5)% = −1.1%, and PIMCO Total Return added value of $R_A = R_P - R_B$ = −0.3% − (0.0%) = −0.3%. These value added numbers represent the skill in security selection within each individual fund.
- Using the actual weights of 68% and 32% in the Fidelity and PIMCO funds, the combined value added from security selection was 0.68(−1.1%) + 0.32(−0.3%) = −0.8%.
- The active asset allocation weights in 2018 were 68% − 60% = +8% for equities and −8% for bonds, so the value added by the active asset allocation decision was 0.08(−4.5%) − 0.08(0.0%) = −0.4%. The total value added by the investor's active asset allocation decision *and* by the mutual funds through security selection was −0.8% − 0.4% = −1.2%. To confirm this total value

added, note that the return on the investor's portfolio was 0.68(−5.6%) + 0.32(−0.3%) = −3.9% and the return on the policy portfolio was 0.60(−4.5%) + 0.40(0.0%) = −2.7%, for a difference of −3.9% − (−2.7) = −1.2%.

Performance attribution systems can be expanded to include several asset classes—for example, stocks, bonds, real estate, and cash (in other words, with $M = 4$ in Equation 4). For a given asset class, the performance attribution system might also include value added from the selection of industries or sectors relative to the benchmark. For example, an equity portfolio might measure value added from over- and underweighting different industry sectors, as well as individual stock selection within those sectors, and a fixed-income portfolio might decompose value added from the mix of sovereign government bonds versus corporate bonds, as well as individual bond selection.

In summary, deviations from portfolio benchmark weights drive the value added by active portfolio management. If every asset in the managed portfolio is held at its benchmark weight, there would be no value added relative to the benchmark. The total value added can be decomposed into various sources that capture the contribution from different decisions, such as asset allocation and security selection.

THE SHARPE RATIO AND THE INFORMATION RATIO 3

☐ calculate and interpret the information ratio (ex post and ex ante) and contrast it to the Sharpe ratio

The risk–return trade-off of a portfolio can be represented in either *absolute* or *relative* terms. The Sharpe ratio provides an absolute expected (*ex ante*) or realized (*ex post*) reward-to-risk measure. As we have noted, however, value added is a relative return comparison. The information ratio provides a benchmark relative expected (*ex ante*) or realized (*ex post*) reward-to-risk measure.

The Sharpe Ratio

The Sharpe ratio is used to compare the portfolio return in excess of a riskless rate with the volatility of the portfolio return. The ratio provides a measure of how much the investor is receiving in excess of a riskless rate for assuming the risk of the portfolio. The Sharpe ratio, SR_P, is calculated for any portfolio, either actively managed or a benchmark, using the formula

$$SR_P = \frac{R_P - R_F}{\sigma_P}, \quad (5)$$

where R_P is the portfolio return, R_F is the risk-free rate, and σ_P is the standard deviation of the portfolio return. In this context, the standard deviation of the portfolio return is often called either volatility or total risk. The Sharpe ratio can be used as an *ex ante* measure of *expected* return and risk, in which case the general formula in Equation 5 would have the expected portfolio return, $E(R_P)$, minus the risk-free rate in the numerator and a forecast of volatility in the denominator. As subjective forecasts, the expected return and standard deviation of return will likely vary among different investors.

The Sharpe ratio can also be used to measure the *ex post* or *realized* performance of a portfolio over some time period. In that case, when applied to multiple time periods, the numerator in Equation 5 is the difference between the average realized portfolio return, $\overline{R_P}$, and the average risk-free rate, $\overline{R_F}$, and the denominator in Equation 5 is the sample standard deviation. The convention for Sharpe ratios is to annualize both the portfolio average return and the portfolio risk. For example, if the past return data are measured monthly, the average monthly return can be multiplied by 12 and the monthly return volatility can be multiplied by the square root of 12. The logic for multiplying the standard deviation by the *square root* of 12 is that variance (i.e., standard deviation squared), under certain assumptions, increases proportionally with time.

Although this scaling convention is common in practice, multiplying monthly returns by a factor of 12 for averages and the square root of 12 for standard deviations ignores the multiplicative (i.e., compound) nature of returns over time. Simple multiplication factors (e.g., 250 and the square root of 250 for annualizing trading-day returns) are only technically correct if the underlying returns are independent and continuously compounded or logarithmic. Similarly, annualized compound returns for the two values in the numerator of the Sharpe ratio (i.e., the portfolio return and the riskless rate) may be used instead of the annualized difference of arithmetic returns. The various methodologies produce slightly different results but should not be a serious problem as long as comparisons between different portfolios use the same approach.

Exhibit 1: Benchmark Sharpe Ratios for 1994–2018 (based on a risk-free rate of 2.3%)

	MSCI World	S&P 500	Russell 2000	MSCI EAFE	Bloomberg Barclays US Aggregate
Average annual return	7.9%	9.9%	10.3%	6.3%	5.0%
Return standard dev.	14.5%	14.4%	19.1%	15.8%	3.5%
Sharpe ratio	0.38	0.53	0.41	0.25	0.77

Exhibit 1 reports the annualized monthly historical return data (not compounded) in US dollars for several different benchmark portfolios for the 25-year period from 1994 to 2018. Long-term *ex post* Sharpe ratios for equity benchmarks have typically fallen within a range of 0.20–0.60, although over a shorter horizon they will vary over a wider range and can be either negative or positive. The Sharpe ratio for the Bloomberg Barclays US Aggregate fixed-income benchmark in Exhibit 1 is particularly high because of the secular decline in interest rates over this 25-year period that boosted the average return for fixed income. Exhibit 2 reports historical return data and Sharpe ratios from 1994 to 2018 for some well-known actively managed mutual funds over the same period. The Sharpe ratios in both exhibits are based on a risk-free rate of 2.3%, the average annualized US Treasury bill return during this 25-year period. The comparison of Sharpe ratios between funds intentionally uses data from the same measurement period. One should not compare the Sharpe ratio of one fund over one period with that of another fund over a different period.

Exhibit 2: Active Fund Sharpe Ratios for 1994–2018 (based on a risk-free rate of 2.3%)

	Fidelity Magellan	Growth Fund of America	Templeton World	T. Rowe Price Small Cap	JPMorgan Bond
Average annual return	8.5%	11.1%	7.9%	11.6%	5.2%
Return standard dev.	16.5%	15.7%	15.2%	16.7%	3.6%
Sharpe ratio	0.38	0.56	0.37	0.56	0.80

Note: The selection of funds for illustration was made without any intended implication, positive or negative, concerning their performance relative to other possible choices.

An important property is that the Sharpe ratio is unaffected by the addition of cash or leverage in a portfolio. Consider a combined portfolio with a weight of w_P on the actively managed portfolio and a weight of $(1 - w_P)$ on risk-free cash. The return on the combined portfolio is $R_C = w_P R_P + (1 - w_P)R_F$, and the volatility of the combined portfolio is just $\sigma_C = w_P \sigma_P$ because the $(1 - w_P)R_F$ portion is risk free. Applying these two relationships in Equation 5 gives the Sharpe ratio for the combined portfolio as

$$\text{SR}_C = \frac{R_C - R_F}{\sigma_C} = \frac{w_P\left(R_P - R_F\right)}{w_P \sigma_P} = \text{SR}_P,$$

which is the same as the Sharpe ratio of the actively managed portfolio. Note that the weight in the combined portfolio, w_P, could be greater than 1, so $(1 - w_P)$ could be negative, indicating that leverage created by *borrowing* risk-free cash and investing in risky assets also does not affect the portfolio's Sharpe ratio.

The proposition that independent of preferences investors should form portfolios using two funds—one of which is the risk-free asset and the other the risky asset portfolio with the highest Sharpe ratio—is known as two-fund separation. On the one hand, if the expected volatility of the risky asset portfolio is higher than the investor prefers, the volatility can be reduced by holding more cash and less of the risky portfolio. On the other hand, if the expected volatility of the risky portfolio is lower than the investor allows, the volatility and expected return can be increased by leverage. For example, suppose an investor believes the performance of the Growth Fund of America shown in Exhibit 2 will repeat going forward but only allows a volatility of 10%. The investor might invest 64% of assets in the Growth Fund of America and 36% in cash to reduce overall portfolio risk. The expected return of the combined portfolio is 0.64(11.1%) + 0.36(2.3%) = 7.9%. The volatility of the combined portfolio is 0.64(15.7%) = 10.0%. The Sharpe ratio of the combined portfolio is (7.9% – 2.3%)/10.0% = 0.56, the same as the 0.56 Sharpe ratio of the Growth Fund of America shown in Exhibit 2.

EXAMPLE 2

Adjusting Risk and Return Using the Sharpe Ratio

Consider an investor choosing between two risky portfolios: a large-cap stock portfolio and a small-cap stock portfolio. Although forecasts about the future are subjective, suppose for simplicity that the investor expects that the future statistics will be those in the following table, with a risk-free rate of 2.3%. The forecasted 0.42 Sharpe ratio of the small-cap portfolio is higher than the 0.40 ratio of the large-cap portfolio, but suppose the investor does not want the high 19.2% volatility associated with the small-cap stocks.

	Large Cap	Small Cap
Expected return	8.2%	10.3%
Expected volatility	14.6%	19.2%
Sharpe ratio	0.40	0.42

1. What percentage of the portfolio would an investor need to hold in cash to reduce the risk of a portfolio invested in the small-cap portfolio and cash to the same risk level as that of the large-cap portfolio?

Solution:

We want to reduce the 19.2% volatility to 14.6% by adding cash. The weight of small-cap stocks in the combined portfolio must therefore be 14.6/19.2 = 76%, leaving a 24% weight in risk-free cash. With that amount of cash, the volatility of the combined portfolio will be 0.76(19.2%) = 14.6%, the same as the large-cap portfolio.

2. Based on your answer to 1, calculate the Sharpe ratio of the small-cap plus cash portfolio.

Solution:

The Sharpe ratio of the combined portfolio is unaffected by the amount in cash, so it remains 0.42.

3. Compare the expected return of the small-cap plus cash portfolio with the expected return of the large-cap portfolio.

Solution:

The expected return of the combined portfolio is 0.76(10.3%) + 0.24(2.3%) = 8.4%, 20 basis points (bps) higher than the 8.2% expected return on the large-cap portfolio, but with the same risk as the large-cap portfolio. To reconfirm, the Sharpe ratio of the combined portfolio is (8.4% – 2.3%)/14.6% = 0.42, the same as the original 0.42 value.

The Information Ratio

The simplest definition of the information ratio compares the active return from a portfolio relative to a benchmark with the volatility of the active return, called “active risk” or “benchmark tracking risk.” The information ratio can be thought of as a way to measure the consistency of active return, as most investors would prefer a more evenly generated value added (low active risk) rather than a lumpy active return pattern. Like the more formal distinction between active portfolio return and alpha, active risk has a more exact beta-adjusted counterpart, which Grinold and Kahn (1999) called “residual risk.” In this discussion, the information ratio is based on the implicit assumption that the beta of the managed portfolio relative to the benchmark is exactly 1.0, although in practice that assumption can be relaxed. For example, Fischer and Wermers (2013) present the information ratio that does not assume beta is 1.

The information ratio tells the investor how much active return has been earned, or is expected to be earned, for incurring the level of active risk. Active return, R_A, is the difference between the managed portfolio return, R_P, and the benchmark portfolio return, R_B. The information ratio of an actively managed portfolio, IR, is calculated by dividing the active return by active risk:

$$\text{IR} = \frac{R_P - R_B}{\sigma\left(R_P - R_B\right)} = \frac{R_A}{\sigma_A} \tag{6}$$

where σ(·) is the standard deviation function—in this case, the standard deviation of the excess return of the portfolio (R_P) over the return of the benchmark (R_B). As with the Sharpe ratio, the typical convention is to annualize both the active return and the active risk. The information ratio can refer to the investor's *ex ante*, or forecasted, active return. Thus, the numerator in Equation 6 would be replaced by the expected returns—that is, $E(R_A) = E(R_P) - E(R_B)$—and the denominator would be the expected active risk. Alternatively, the calculation of an *ex post*, or historical, information ratio would use realized average active returns and the realized sample standard deviation of the active return.

Two investment strategies and associated terminology can help reinforce the conceptual distinction between the Sharpe ratio and the information ratio. First, a "closet index fund" (a fund that advertises itself as being actively managed but is actually close to being an index fund) will have a Sharpe ratio that is close to the benchmark because the excess return and volatility will be similar to the benchmark. However, the closet index fund will have a small amount of active risk, although positive by definition like any volatility estimate. While there may be little active risk, the information ratio of a closet index fund will likely be close to zero or slightly negative if value added cannot overcome the management fees. If one has the actual holdings of the fund, closet indexing is easy to detect on the basis of a measurement called "active share," a measure of how similar a portfolio is to its benchmark. [Cremers and Petajisto (2009) defined active share as half the sum of the absolute values of the active weights.] As a second example, the Sharpe ratio and the information ratio for a market-neutral long–short equity fund (a fund with offsetting long and short positions that has a beta of zero with respect to the market) would be identical if we consider the benchmark to be the riskless rate because the excess return and active return would be the same calculation, as would be total risk and active risk.

Exhibit 3 shows historical information ratios for the mutual funds in Exhibit 2, with the benchmark portfolio for each calculation shown at the bottom of Exhibit 3. The average active return in the first row of Exhibit 3 can be calculated by subtracting the specified benchmark average return in Exhibit 1 from the average fund return in Exhibit 2. The active risk is the annualized standard deviation of the return differences from 1994 to 2018, which cannot be verified with just the summary data in Exhibit 1 and Exhibit 2.

As shown in Exhibit 3, *ex post* information ratios will be negative if the active return is negative. In fact, under the zero-sum property of active management, the average realized information ratio across investment funds with the same benchmark should be about zero. The realized information ratios in Exhibit 3 are within a range of about −0.30 to +0.30, although the range would be much wider over shorter periods. Of course, *ex ante*, or before the fact, if an investor did not expect the information ratio to be positive, he or she would simply invest in the benchmark. Note that ranking by active risk, a relative measure, does not necessarily equate to ranking by total risk, an absolute measure. For example, the relative risk of Fidelity Magellan in Exhibit 3 is slightly lower than the relative risk of the Growth Fund of America; however, the absolute risk of Fidelity Magellan in Exhibit 2 is slightly higher.

Exhibit 3: Active Fund Information Ratios for 2014–2018

	Fidelity Magellan	Growth Fund of America	Templeton World	T. Rowe Price Small Cap	JPMorgan Bond
Active return	−1.4%	1.2%	0.0%	1.4%	0.2%
Active risk	5.1%	6.2%	5.0%	4.7%	1.0%
Information ratio	−0.27	0.20	0.00	0.29	0.19
Benchmark	S&P 500	S&P 500	MSCI World	Russell 2000	Bloomberg Barclays US Aggregate

Unlike the Sharpe ratio, the information ratio is affected by the addition of cash or the use of leverage. For example, if the investor adds cash to a portfolio of risky assets, the information ratio for the combined portfolio will generally shrink. However, the information ratio of an unconstrained portfolio is unaffected by the aggressiveness of active weights. Specifically, if the active security weights, Δw_i, defined as deviations from the benchmark portfolio weights, are all multiplied by some constant, *c*, the information ratio of an actively managed portfolio will remain unchanged.

Recall the expression for the active return of a managed portfolio in Equation 3. If each active weight in Equation 3 is multiplied by some constant, *c*, then the active return on the altered portfolio, R_C, is

$$R_C = \sum_{i=1}^{N} c\Delta w_i R_{Ai} = c\sum_{i=1}^{N} \Delta w_i R_{Ai} = cR_A.$$

Similarly, the active risk of the altered portfolio is $c\sigma_A$, so the information ratio of the altered portfolio is

$$\text{IR}_C = \frac{cR_A}{c\sigma_A} = \text{IR},$$

the same as that of the actively managed portfolio with no proportional increase in the active weights. Specifically, if the active weights in a managed portfolio are all doubled, the expected active return (or realized average active return) would be doubled, along with the expected or realized active risk, leaving the information ratio unchanged.

Of course, an outside investor would not be able to adjust the active risk of an existing fund by changing the individual asset active weight positions, but the same objective can be met by taking positions in the benchmark portfolio. For example, if the active risk of a fund is 5.0%, combining that fund in an 80/20 mix with the benchmark portfolio (i.e., a benchmark portfolio weight of 20%) will result in an active risk of the combined portfolio of 0.80(5.0%) = 4.0%, with a proportional reduction in the active return. Similarly, the investor can short sell the benchmark portfolio and use the proceeds to invest in the actively managed fund to increase the active risk and return. Note that in practice, institutional investors might simply reduce the amount they would have otherwise invested in the benchmark portfolio—or, if possible, another actively managed fund—rather than employ an explicit short sell.

CONSTRUCTING OPTIMAL PORTFOLIOS

4

☐ calculate and interpret the information ratio (ex post and ex ante) and contrast it to the Sharpe ratio

An important concept from basic portfolio theory is that with a risk-free asset, the portfolio on the efficient frontier of risky assets that is tangent to a ray extended from the risk-free rate is optimal because it has the highest possible Sharpe ratio. Thus, given the opportunity to adjust absolute risk and return with cash or leverage, the overriding objective is to find the single risky asset portfolio with the maximum Sharpe ratio, whatever the investor's risk aversion. A similarly important property in active management theory is that given the opportunity to adjust active risk and return by investing in both the actively managed and benchmark portfolios, the squared Sharpe ratio of an actively managed portfolio is equal to the squared Sharpe ratio of the benchmark plus the information ratio squared:

$$\mathrm{SR}_P^2 = \mathrm{SR}_B^2 + \mathrm{IR}^2. \tag{7}$$

Equation 7 implies that the active portfolio with the highest (squared) information ratio will also have the highest (squared) Sharpe ratio. (Note that Equation 7 is not practical for comparisons of investment skill involving negative IR because the sign is lost in squaring.) As a consequence, according to mean−variance theory, the expected information ratio is the single best criterion for assessing active performance among various actively managed funds with the same benchmark (Grinold 1989). For any given asset class, an investor should choose the manager with the highest expected skill as measured by the information ratio, because investing with the highest information-ratio manager will produce the highest Sharpe ratio for the investor's portfolio.

The preceding discussion on adjusting active risk raises the issue of determining the *optimal* amount of active risk, without resorting to utility functions that measure risk aversion. For unconstrained portfolios, the level of active risk that leads to the optimal result in Equation 7 is

$$\sigma\left(R_A\right) = \frac{\mathrm{IR}}{\mathrm{SR}_B}\sigma_B, \tag{8}$$

where σ_B is the standard deviation of the benchmark return. (Note that the right-hand side of the equation should be multiplied by the benchmark beta of the actively managed portfolio if that value is different from 1.) This Sharpe ratio-maximizing level of active risk or "aggressiveness" comes from the general mean–variance optimality condition that the ratio of expected active return to active return variance of the managed portfolio be set equal to the ratio of expected benchmark excess return to benchmark return variance:

$$\frac{E\left(R_A\right)}{\sigma_A^2} = \frac{E\left(R_B - R_F\right)}{\sigma_B^2}.$$

For example, if the actively managed portfolio has an information ratio of 0.30 and active risk of 8.0% and the benchmark portfolio has an expected excess return of 6.4% and total risk of 16.0% resulting in a Sharpe ratio of 0.40, then according to Equation 8, the optimal amount of aggressiveness in the actively managed portfolio is (0.30/0.40)16.0% = 12.0%. If the actively managed portfolio is constructed with this amount of active risk, the Sharpe ratio will be $(0.40^2 + 0.30^2)^{1/2} = 0.50$, as shown in Equation 7. To verify this Sharpe ratio, note that the more aggressively managed portfolio in this example has an expected active return of (0.30)12.0% = 3.6% over the benchmark, or a total expected excess return of 6.4% + 3.6% = 10.0%. By definition, the total risk of the actively managed portfolio is the sum of the benchmark return variance and active return variance,

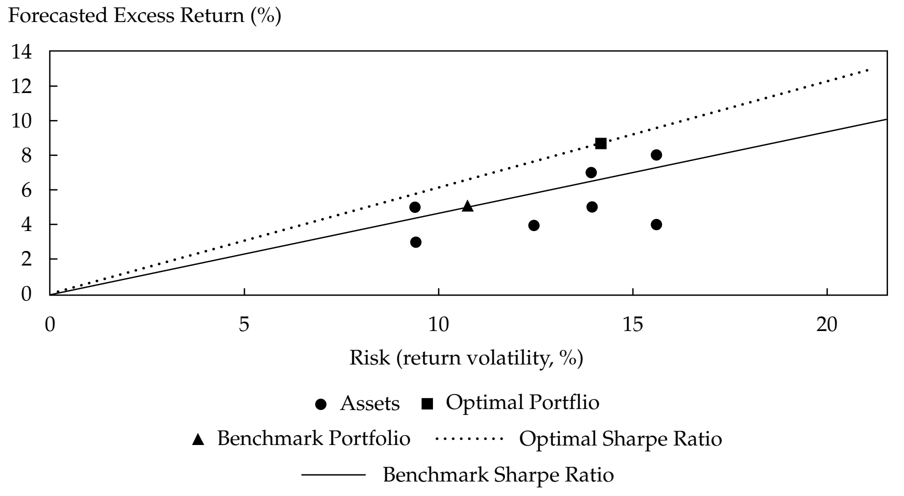
$$\sigma_P^2 = \sigma_B^2 + \sigma_A^2.$$

At the optimal active risk of 12.0%, the total portfolio risk is $(16.0^2 + 12.0^2)^{1/2}$ = 20.0%, verifying the maximum possible Sharpe ratio of 10.0/20.0 = 0.50.

The initial actively managed portfolio has active risk of only 8.0%, whereas the optimal amount required under the assumed information ratio needed to maximize the Sharpe ratio is 12.0%. The actively managed portfolio would thus need to be managed more aggressively to increase the active risk while preserving the same information ratio; alternatively, the investor could short the benchmark and use the proceeds to increase the amount invested in the actively managed fund. The proportion required to be invested in the actively managed fund would be 12.0/8.0 = 1.5 times; shorting the benchmark by 0.5 times would fund the increase.

For readers familiar with risk–return charts in basic portfolio theory, Exhibit 4 will help illustrate these concepts.

Exhibit 4: Portfolio Risk and Return



Several individual risky assets are plotted in Exhibit 4 in terms of their forecasted return in excess of the risk-free rate ("excess return") on the vertical axis and risk on the horizontal axis. The values for the individual assets are based on subjective assessments supplied by the investor. The theory described here explains how to optimally employ those expectations assuming they are based on reasonable judgment. Using the benchmark portfolio weights (not shown), the risks and expected returns of the individual assets combine into the benchmark portfolio risk and expected return shown in Exhibit 4. Because the expected returns plotted along the vertical axis are in excess of the risk-free rate, the slope of a line that emanates from the origin (zero risk and zero excess return) is the Sharpe ratio of the benchmark portfolio. Specifically, the Sharpe ratio of the benchmark portfolio (i.e., slope of the dark line) in Exhibit 4 is the expected excess return of 5.0% divided by return volatility of 10.8%, or 5.0%/10.8% = 0.46.

Because of diversification, the Sharpe ratio of the benchmark portfolio is higher than those of most of the individual assets; however, the benchmark portfolio does not have the highest possible Sharpe ratio of all portfolios that can be constructed from these assets. In fact, the optimal portfolio (i.e., mean–variance efficient frontier portfolio with the highest possible Sharpe ratio) shown in Exhibit 4 has an expected excess return of 8.7% and return volatility of 14.2%, resulting in a Sharpe ratio of 8.7%/14.2% = 0.61 (i.e., slope of the dotted line). This higher Sharpe ratio could be retained in the portfolio while adjusting the level of risk through the use of cash or leverage. For example, the risk of the optimal portfolio could be reduced along the dotted line to the benchmark portfolio risk of 10.8% with an expected excess return of 0.61(10.8%) = 6.6%, compared with the benchmark expected excess return of 5.0%.

Exhibit 5: Portfolio Active Risk and Return

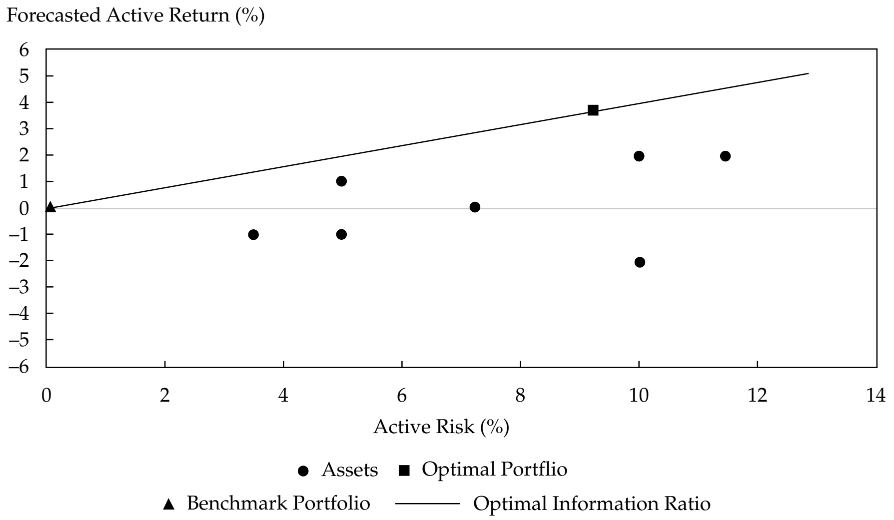


Exhibit 5 plots the same individual assets, the benchmark portfolio, and the optimal portfolio from Exhibit 4 in terms of their expected *active* return on the vertical axis and *active* risk on the horizontal axis. By definition, the benchmark portfolio is plotted at the origin in Exhibit 5 with zero active return and zero active risk. The individual assets have both positive and negative expected active returns compared with the benchmark portfolio return, whereas the optimal portfolio has a positive active return of 3.8% and active risk of 9.4%. The information ratio of the optimal portfolio is therefore 3.8%/9.4% = 0.40, the slope of the dark line in Exhibit 5. The information ratio of the optimal portfolio is higher than that of any of the individual assets; in fact, it can be shown to be the square root of the sum of the squared values of the individual assets' information ratios, similar to Equation 7, including those assets with negative information ratios. The asset weights required for the construction of this optimal portfolio are the subject of the next section, but here we note they might be negative for negative-IR assets—that is, short sells of individual assets may be required.

Although the information ratio will remain constant at 0.40, various levels of aggressiveness can be applied to the actively managed portfolio in Exhibit 5 by scaling the optimal active weights or, alternatively, taking a position in the benchmark portfolio—leading to portfolios that plot along the dark line. But to construct the *optimal* actively managed portfolio given the assumed information ratio, the active risk must be adjusted to a level of (0.40/0.46)10.8% = 9.4% in accordance with Equation 8. Specifically, this level of aggressiveness is required to construct the optimal portfolio in Exhibit 4, and according to Equation 7, the Sharpe ratio of this optimal portfolio is $(0.46^2 + 0.40^2)^{1/2}$ = 0.61.

As we will see later, optimal levels of active risk in equity management practice are typically lower than those shown in this numerical example because the underlying portfolios are constrained to be long only, leading to information ratios that are substantially lower. As the information ratio gets close to zero, either because of constraints or because the manager is judged to be less skilled, the optimal amount of active risk in Equation 8 goes to zero (i.e., the optimal portfolio becomes the passive benchmark portfolio).

EXAMPLE 3

Expected Value Added Based on the Information Ratio

Suppose that the historical performance of the Fidelity Magellan and Growth Fund of America mutual funds from Exhibits 2 and 3 are indicative of the future performance of hypothetical funds, "Fund I" and "Fund II." In addition, suppose that the historical performance of the S&P 500 benchmark portfolio shown in Exhibit 1 is indicative of expected returns and risk going forward, as shown in the following excerpts. We use historical values in this problem for convenience, but in practice the forecasted, or expected, values for both the benchmark portfolio and the active funds would be subjectively determined by the investor.

Excerpted from and (based on a risk-free rate of 2.3%)

	S&P 500	Fidelity Magellan (Fund I)	Growth Fund of America (Fund II)
Average annual return	9.9%	8.5%	11.1%
Return standard dev.	14.4%	16.5%	15.7%
Sharpe ratio	0.53	0.38	0.56

Excerpted from

	Fidelity Magellan (Fund I)	Growth Fund of America (Fund II)
Active return	−1.4%	1.2%
Active risk	5.1%	6.2%
Information ratio	−0.27	0.20
Benchmark	S&P 500	S&P 500

1. State which of the two actively managed funds, Fund I or Fund II, would be better to combine with the passive benchmark portfolio and why.

Solution:

Fund II is better, as measured by the combined Sharpe ratio, because Fund II has the higher expected information ratio: 0.20 compared with −0.27 in Fund I.

2. Calculate the possible improvement over the S&P 500 Sharpe ratio from the optimal deployment of Fund II, which has an expected information ratio of 0.20.

Solution:

Properly combined with the S&P 500 benchmark portfolio, Fund II has the potential to increase the expected Sharpe ratio from 0.53 for the passive benchmark portfolio to an expected Sharpe ratio of $(0.53^2 + 0.20^2)^{1/2} = 0.57$.

3. Fund I comes with an active (i.e., benchmark relative) risk of 5.1%, but the investor wants to adjust the active risk to 5.4%. Describe how that adjustment would be made. (No calculations are required; give a qualitative description.)

Solution:

To increase the active risk of Fund I, the investor would need to be more aggressive in managing the portfolio, take a short (i.e., negative) position in the benchmark, or, more simply, invest less than he or she otherwise would have in the benchmark or another actively managed fund.

4. Fund II comes with an active risk of 6.2%. Determine the weight of the benchmark portfolio required to create a combined portfolio with the highest possible expected Sharpe ratio.

Solution:

According to Equation 8, the optimal amount of active risk is (0.20/0.53)14.4% = 5.4%. A positive position in the benchmark is needed to adjust the active weight down from 6.2%. Specifically, the benchmark portfolio weight needed to adjust the active risk in Fund II is 1 − 5.4%/6.2% = 13%.

Note that at the 5.4% optimal level of active risk, Fund II has an expected active return of 0.20(5.4%) = 1.1%, a total expected excess return of 7.6% + 1.1% = 8.7%, and a total risk of $(14.4^2 + 5.4^2)^{1/2} = 15.4\%$. The result is an expected Sharpe ratio of 8.7/15.4 = 0.57, the same as the value calculated for Question 2.

In summary, the information ratio is active return over active risk (in contrast to the excess return-to-risk measure known as the Sharpe ratio). Information ratios help investors focus on the relative valued added by active management. The information ratio is unaffected by the aggressiveness of the active weights (i.e., deviations from benchmark weights) in the managed portfolio because both the active return and the active risk increase proportionally. The potential improvement in an active portfolio's expected Sharpe ratio compared with the benchmark's Sharpe ratio is a function of the squared information ratio. Thus, the expected information ratio becomes the single best criterion for constructing an actively managed portfolio, and the *ex post* information ratio is the best criterion for evaluating the past performance of various actively managed funds.

5 ACTIVE SECURITY RETURNS AND THE FUNDAMENTAL LAW OF ACTIVE MANAGEMENT

- ☐ describe and interpret the fundamental law of active portfolio management, including its component terms—transfer coefficient, information coefficient, breadth, and active risk (aggressiveness)
- ☐ explain how the information ratio may be useful in investment manager selection and choosing the level of active portfolio risk

The fundamental law is a framework for thinking about the potential value added through active portfolio management. The framework can be used to size individual asset active weights, estimate the expected value added of an active management strategy, or measure the realized value added after the fact; however, the most common use is the description and evaluation of active management strategies. The law itself is a mathematical relationship that relates the expected information ratio of an actively managed portfolio to a few key parameters.

Active Security Returns

On the basis of the prior section, we assume that the investor is concerned about maximizing the managed portfolio's active return subject to a limit on active risk (also called "benchmark tracking risk"). To this end, the investor uses forecasts for each security of the active return, R_{Ai}, or thus the benchmark relative return,

$$R_{Ai} = R_i - R_B, \tag{9}$$

for the N individual assets that might be included in the portfolio. Our notation for the investor's forecasts of the active security returns is μ_i (Greek letter mu). The term μ_i can be thought of as the security's expected active return, $\mu_i = E(R_{Ai})$, referring to the investor's subjective expectation, in contrast to an expectation based on a formal equilibrium model.

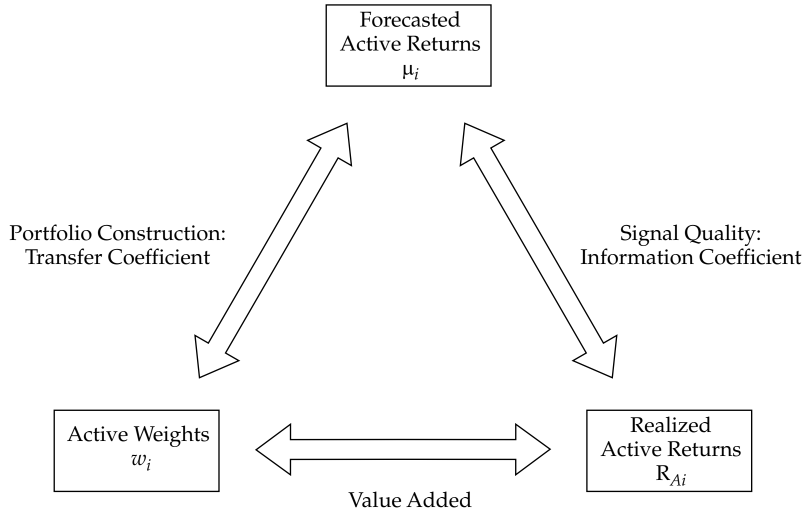
Although we focus on the simple definition of active security return in Equation 9, there are several possible choices depending on the assumed risk model (i.e., statistical model of returns) and the desired trade-off between a conceptual treatment and more complex but implementable formulas. For example, Equation 9 can be modified to define the active security return as the residual return in a single-factor model, $R_{Ai} = R_i - \beta_i R_B$, where β_i is the sensitivity of the security return to the benchmark return. Although this expression may appear to be related to the CAPM, the benchmark return may or may not be the market return. Moreover, the fundamental law does not require the empirical validity of the CAPM, the multi-factor APT (arbitrage pricing theory), or any other equilibrium theory of required returns. The individual security active return can also be defined as the residual return in a multi-factor model:

$$R_{Ai} = R_i - \sum_{j=1}^{K} \beta_{j,i} R_j,$$

with K market-wide factor returns, R_j, and security sensitivities, $\beta_{j,i}$, to those factors.

Exhibit 6 provides a conceptual diagram in which to think about the various parameters in the fundamental law of active management. At the three corners of the triangle are the sets of forecasted active returns, μ_i, active portfolio weights, Δw_i, and realized active returns, R_{Ai}. The base of the triangle reflects the realized value added through active management, defined as the difference between the realized returns on the actively managed portfolio and the benchmark portfolio. Value added is the sum of the products of active weights and active returns for the i = 1 to N securities in the portfolio, as shown in Equation 3. The value of this sum is ultimately a function of the correlation coefficient between the active weights, Δw_i, and realized active returns, R_{Ai}.

Exhibit 6: The Correlation Triangle



To understand the role of the correlation coefficient, consider the following algebraic expansion of Equation 3 that uses COV, STD (σ), and COR (ρ) to designate the covariance, standard deviation, and correlation coefficient functions, respectively:

$$R_A = \sum_{i=1}^{N} \Delta w_i R_{Ai}$$
$$= \text{COV}(\Delta w_i, R_{Ai})N$$
$$= \rho(\Delta w_i, R_{Ai})\,\sigma(\Delta w_i)\,\sigma(R_{Ai})N$$

The exact equalities in this expansion depend on the fact that the cross-sectional means of active weights and active returns are zero. Specifically, the population covariance between two variables, X and Y, is calculated as $\text{COV}(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(X_i - \overline{X})(Y_i - \overline{Y})$, or simply $\text{COV}(X,Y) = \frac{1}{N}\sum_{i=1}^{N}X_i Y_i$.

Similarly, the population variance for a single variable is $\text{VAR}(X) = \frac{1}{N}\sum_{i=1}^{N}X_i^2$ if the mean is zero.

In Exhibit 6, the arrows in the legs of the triangle represent the correlation between the quantities at the corners of the triangle to which the arrows point. While the arrow at the base of the triangle reflects value added, a better understanding of the sources and limitations of value added can be obtained by examining the correlations on the two vertical legs. First, there is little hope of adding value if the investor's forecasts of active returns do not correspond at least loosely to the realized active returns. Signal quality is measured by the correlation between the forecasted active returns, μ_i, at the top of the triangle, and the realized active returns, R_{Ai}, at the right corner, commonly called the information coefficient (IC). Investors with higher IC, or ability to forecast returns, will add more value over time but only to the extent that those forecasts are exploited in the construction of the managed portfolio. The correlation between any set of active weights—Δw_i, in the left corner, and forecasted active returns, μ_i, at the top of the triangle—measures the degree to which the investor's forecasts are translated into active weights, called the transfer coefficient (TC).

The mathematics of the fundamental law were introduced by Grinold (1989) and further developed by Clarke, de Silva, and Thorley (2002). The mean–variance-optimal active security weights for uncorrelated active returns, subject to a limit on active portfolio risk, are given by

$$\Delta w_i^* = \frac{\mu_i}{\sigma_i^2} \frac{\sigma_A}{\sqrt{\sum_{i=1}^{N} \frac{\mu_i^2}{\sigma_i^2}}},$$

where σ_A represents active portfolio risk and σ_i is the forecasted volatility of the active return on security *i*. This formula for active weights makes intuitive sense. The deviation (positive or negative) from the benchmark weight for security *i* are higher for larger values of the forecasted active return, μ_i, but are reduced by forecasted volatility, σ_i. In addition, the active weights are scaled by the active risk of the portfolio, σ_A, so that the desire for more active portfolio risk requires larger individual active weights.

In addition to employing mean–variance optimization, proofs of the fundamental law generally assume that active return forecasts are scaled prior to optimization using the Grinold (1994) rule:

$$\mu_i = \text{IC}\sigma_i S_i, \quad (10)$$

where IC is the expected information coefficient and S_i represents a set of standardized forecasts of expected returns across securities, sometimes called "scores." Scores with a cross-sectional variance of 1 are used in Equation 10 to ensure that the scaling process using the multipliers σ_i (separate values for individual securities) and IC (one value for all securities) result in expected active returns of the correct magnitude. Specifically, if the assumed IC value is low, then the cross-sectional variation of the expected active returns in Equation 10 will be low. However, the exact process for calculating expected active returns may be more involved than the simple rule indicated in Equation 10 depending on how the investor's views on individual asset returns are originally formulated.

Using the Grinold rule shown in Equation 10, the mean–variance optimal active weights are

$$\Delta w_i^* = \frac{\mu_i}{\sigma_i^2} \frac{\sigma_A}{\text{IC}\sqrt{\text{BR}}}, \quad (11)$$

where IC stands for information coefficient and BR, which has replaced the symbol *N*, stands for breadth. We discuss these two key fundamental law parameters next.

As previously stated, IC is the *ex ante* (i.e., anticipated) cross-sectional correlation between the *N* forecasted active returns, μ_i, and the *N* realized active returns, R_{Ai}. To be more accurate, IC is the *ex anterisk-weighted* correlation

$$\text{IC} = \rho\left(\frac{R_{Ai}}{\sigma_i}, \frac{\mu_i}{\sigma_i}\right), \quad (12)$$

where ρ(·) indicates correlation. As a correlation coefficient, IC can take on values anywhere from −1.00 to +1.00, although small positive values less than 0.20 are often the norm. The *ex ante*, or anticipated, IC must be positive or the investor would not pursue active management but simply invest in the passive benchmark. Later, we will discuss the realized, or *ex post*, information coefficient in terms of measuring active management performance after the fact—where the realized information coefficient might be either positive or negative, leading to positive or negative value added.

The other important fundamental law parameter in Equation 11 is BR, or breadth, conceptually equal to the number of independent decisions made per year by the investor in constructing the portfolio. The simplest case for the calculation of breadth is a single-factor risk model, where the only source of correlation between the securities is the common market factor and decisions about the active return for any given security are independent from one year to the next. In this case, breadth is equal to the number of securities: Each active return is independent from the other active return forecasts for that period and independent from the forecast for that security in subsequent periods.

However, most risk models will incorporate other factors—for example, economic sectors or industries. If the risk model includes the assumption that all the securities within a given industry are positively correlated, then part of the forecast that the active returns for securities in that industry will be higher or lower is based on just one perspective the investor has about the industry. In this case, breadth is intuitively lower than the number of securities. Alternatively, breadth can be higher than the number of securities if factors in the risk model suggest that their active returns are negatively correlated. For these more complicated cases, breadth will be a non-integer number, as noted in Clarke, de Silva, and Thorley (2006).

Similarly, if some aspect of a security is fairly constant over time and the investor makes decisions about expected active return based on that characteristic, then breadth over time is lower. Alternatively, if the investor makes quarterly or monthly forecasts about a security that are truly independent over time, then breadth can be as high as the number of securities times the number of rebalancing periods per year.

EXAMPLE 4

Scaling Active Return Forecasts and Sizing Active Weights

Consider the simple case of four individual securities whose active returns are assumed to be uncorrelated with each other and have active return volatilities of 25.0% and 50.0%. After some analysis, an active investor believes the first two securities will outperform the other two over the next year and thus assigns scores of +1 and −1 to the first and second groups, respectively. The scenario is depicted in the following exhibit:

Security	Score	Volatility
#1	1.0	25.0%
#2	1.0	50.0%
#3	−1.0	25.0%
#4	−1.0	50.0%

1. Assume that the anticipated accuracy of the investor's ranking of securities is measured by an information coefficient of IC = 0.20. What are the forecasted active returns for each of the four securities using the scaling rule $\mu_i = \text{IC}\sigma_i S_i$?

Solution:

The forecasted active return to Security #1 is 0.20(25.0%)(1.0) = 5.0%. Similar calculations for the other three securities are shown in the following exhibit.

Security	Score	Active Return Volatility	Expected Active Return
#1	1.0	25.0%	5.0%
#2	1.0	50.0%	10.0%
#3	−1.0	25.0%	−5.0%
#4	−1.0	50.0%	−10.0%

2. Given the assumptions that the four securities' active returns are uncorrelated with each other and forecasts are independent from year to year, what is the breadth of the investor's forecasts?

Solution:

If the active returns are uncorrelated with each other and the forecasts are independent from year to year, then the investor has made four separate decisions and breadth is BR = 4, the number of securities.

3. Suppose the investor wants to maximize the expected active return of the portfolio subject to an active risk constraint of 9.0%. Calculate the active weights that should be assigned to each of these securities using the formula $\Delta w_i^* = \frac{\mu_i}{\sigma_i^2} \frac{\sigma_A}{\text{IC}\sqrt{\text{BR}}}$.

Solution:

The size of the active weight for Security #1 is $\Delta w_i^* = \frac{0.05}{0.25^2} \frac{0.09}{0.20\sqrt{4}} = 18\%$.

Similar calculations for the other four securities are shown in the following exhibit.

Security	Expected Active Return	Active Return Volatility	Active Weight
#1	5.0%	25.0%	18.0%
#2	10.0%	50.0%	9.0%
#3	−5.0%	25.0%	−18.0%
#4	−10.0%	50.0%	−9.0%

The Basic Fundamental Law

On the basis of Equation 3, the anticipated value added for an actively managed portfolio, or expected active portfolio return, is the sum product of active security weights and forecasted active security returns:

$$\text{E}\left(R_A\right) = \sum_{i=1}^{N} \Delta w_i \mu_i.$$

Using the optimal active weights in Equation 11 and forecasted active security returns in Equation 10, the expected active portfolio return is

$$\text{E}\left(R_A\right)^* = \text{IC}\sqrt{\text{BR}}\,\sigma_A, \quad (13)$$

where the * indicates that the actively managed portfolio is constructed from *optimal* active security weights, Δw_i^*. Remember that the algebra for this result assumes that breadth is the number of securities: BR = N. A more general proof where breadth is different from the number of securities is provided in Clarke, de Silva, and Thorley (2006).

The basic fundamental law of active management in Equation 13 states that the optimal expected active return, $\text{E}(R_A)^*$, is the product of three key parameters: the assumed information coefficient, IC, the square root of breadth, BR, and portfolio active risk, σ_A. Using Equation 13, we can also express the information ratio of the unconstrained optimal portfolio, $\text{E}(R_A)^*/\sigma_A$, as the product of just two terms: $\text{IR}^* = \text{IC}\sqrt{\text{BR}}$.

EXAMPLE 5

The Basic Fundamental Law

Consider the simple case of four individual securities whose active returns are uncorrelated with each other and forecasts are independent from year to year. The active return forecasts, active risks, and the active weights for each security are shown in the following exhibit.

Security	Expected Active Return	Active Return Volatility	Active Weight
#1	5.0%	25.0%	18%
#2	10.0%	50.0%	9%
#3	−5.0%	25.0%	−18%
#4	−10.0%	50.0%	−9%

1. Suppose that the benchmark portfolio for these four securities is equally weighted (i.e., $w_{B,i}$ = 25% for each security) and that the forecasted return on the benchmark portfolio is 10.0%. What are the portfolio weights and the total expected returns for each of the four securities?

Solution:

The portfolio weight for Security #1 is the benchmark weight plus the active weight, 25% + 18% = 43%. The total expected return for Security #1 is the expected benchmark return plus the expected active return, 10.0% + 5.0% = 15.0%. Similar calculations for the other three securities are shown in the following exhibit.

Security	Total Weight	Total Return Forecast
#1	43%	15.0%
#2	34%	20.0%
#3	7%	5.0%
#4	16%	0.0%
	100%	

2. Calculate the forecasted total return and active return of the managed portfolio.

Solution:

The forecasted total return of the portfolio is the sum of portfolio weights times total returns for each security: 0.43(15.0) + 0.34(20.0) + 0.07(5.0) + 0.16(0.0) = 13.6%. The expected active return of the portfolio is the managed portfolio return minus the benchmark return: 13.6 − 10.0 = 3.6%. Alternatively, the calculation is the sum of active weights times active returns for each security: 0.18(5.0%) + 0.09(10.0%) − 0.18(−5.0%) − 0.09(−10.0%) = 3.6%.

3. Calculate the active risk of the managed portfolio.

Solution:

The active risk of the managed portfolio is the square root of the sum of active weights squared times the active volatility squared for each security, which gives $[0.18^2 \times 25.0^2 + 0.09^2 \times 50.0^2 + (-0.18)^2 \times 25.0^2 + (-0.09)^2 \times 50.0^2]^{1/2} = 9.0\%$.

4. Verify the basic fundamental law of active management using the expected active return and active risk of the managed portfolio. The individual security active return forecasts and active weights were sized using an information coefficient of IC = 0.20, breadth of BR = 4, and active risk of $\sigma_A = 9.0\%$.

Solution:

The basic fundamental law states that the expected active portfolio return is $IC\sqrt{BR}\,\sigma_A = 0.20 \times 4^{1/2} \times 9.0 = 3.6\%$, which is consistent with the calculation in the Solution to 2. Alternatively, the information ratio of 3.6/9.0 = 0.40 confirms the basic fundamental law that $IR^* = IC\sqrt{BR} = 0.20 \times 4^{1/2} = 0.40$.

6 THE FULL FUNDAMENTAL LAW

- ☐ describe and interpret the fundamental law of active portfolio management, including its component terms—transfer coefficient, information coefficient, breadth, and active risk (aggressiveness)
- ☐ explain how the information ratio may be useful in investment manager selection and choosing the level of active portfolio risk

Although we were able to derive an analytic (i.e., formula-based) solution for the set of unconstrained optimal active weights in Equation 11, a number of practical or strategic constraints are often imposed in practice. For example, if the unconstrained active weight of a particular security is negative and large, that might lead to a negative absolute weight or short sell of the security. Many investors are constrained to be long only, either by regulation or by preference because of the extra complexity and costs of short selling. For quantitatively oriented investors, optimal solutions for active weights under long-only constraints, limits on turnover, ESG screens, or other constraints generally require the use of a numerical optimizer. Alternatively, one can use the fundamental law framework to better analyze the active weights that are subjectively determined by less quantitative techniques.

Let Δw_i (without an *) represent the *actual* active security weights for a constrained portfolio—in contrast to the optimal active weights, Δw_i^*, specified in Equation 11. As explained previously, the transfer coefficient, TC, is essentially the cross-sectional correlation between the forecasted active security returns and actual active weights. To be more precise, for a single-factor risk model, TC is the following *risk-weighted* correlation:

$$TC = \rho(\mu_i/\sigma_i, \Delta w_i \sigma_i).$$

Based on the correspondence between optimal active weights and forecasted active returns in Equation 11, the transfer coefficient can also be expressed as the risk-weighted correlation between the optimal active weights and the actual active weights, $TC = \rho\left(\Delta w_i^* \sigma_i, \Delta w_i \sigma_i\right)$.

As a correlation coefficient, TC can take on values anywhere from −1.00 to +1.00, although TC values are typically positive and range from about 0.20 to 0.90. A low TC results from the formal or informal constraints imposed on the structure of the portfolio. In fact, at TC = 0.00, there would be no correspondence between the active return forecasts and active weights taken and thus no expectation of value added from active management. In contrast, TC = 1.00 (no binding constraints) represents a perfect correspondence between active weights taken and forecasted active returns, allowing the full expected value added to be reflected in the portfolio structure. The portfolio TC could even conceivably be negative if relative weights are negatively correlated with current expected returns because the portfolio needs rebalancing.

Including the impact of the transfer coefficient, the expanded fundamental law is expressed in the following equation:

$$\mathrm{E}\left(R_A\right) = (\mathrm{TC})\ (\mathrm{IC})\sqrt{\mathrm{BR}}\,\sigma_A, \quad (14)$$

where an * is not used because the managed portfolio is constructed from *constrained* active security weights, Δw_i. The expanded fundamental law of active management shown in Equation 14, which we will refer to simply as the *fundamental law* henceforward, states that the expected active return, $\mathrm{E}(R_A)$, is the product of four key parameters: the transfer coefficient, TC, the assumed information coefficient, IC, the square root of breadth, BR, and portfolio active risk, σ_A. Using Equation 14, we can also express the portfolio's information ratio, $\mathrm{E}(R_A)/\sigma_A$, as the product of just three terms: $\mathrm{IR} = (\mathrm{TC})\ (\mathrm{IC})\sqrt{\mathrm{BR}}$.

The fundamental law as stated in Equation 14, although more practical than Equation 13, is still based on a simple risk model for the individual securities. Specifically, the equations in this section are based on the simplifying assumption of a single-index model; so, the active security returns are residual returns and are uncorrelated with each other. If we go even further in terms of simplicity and assume that the individual securities all have the same residual volatility, then the correlation formulas for IC and TC do not need to be risk weighted. Alternatively, we could move in the direction of more complexity by using the single-factor risk model with factor sensitivity: $R_{Ai} = R_i - \beta_i R_B$. In quantitative portfolio management practice, even more sophisticated multi-factor risk models are used with correspondingly more complex fundamental law parameter values, although the basic form of Equation 14 is preserved.

EXAMPLE 6

The Expanded Fundamental Law

Consider the simple case of four individual securities whose active returns are uncorrelated with each other and forecasts are independent from year to year. The securities have a range of active return forecasts, risks, optimal active weights, and actual active weights as given in the following exhibit. The optimal active weights are based on a formula for maximizing the active return of a managed portfolio for a given level of active risk. The actual active weights are the result of a numerical optimizer with a number of constraints, in addition to the active risk constraint of 9.0%.

Security	Expected Active Return	Active Return Volatility	Optimal Active Weight	Actual Active Weight
#1	5.0%	25.0%	18%	6%
#2	10.0%	50.0%	9%	4%
#3	−5.0%	25.0%	−18%	7%
#4	−10.0%	50.0%	−9%	−17%

1. Calculate the transfer coefficient (TC) as the risk-weighted correlation coefficient between the four active return forecasts and the four actual active

weights. Compare this number with the transfer coefficient for the optimal active weights.

Solution:

The transfer coefficient is the correlation between the risk-weighted expected active returns and actual active weights: TC = $\rho(\Delta w_i \sigma_i, \mu_i/\sigma_i)$, where ρ denotes correlation (you can use the Microsoft Excel function CORREL) with four pairs of numbers. The risk-weighted values for Security #1 are $\Delta w_1 \sigma_1$ = 0.06(25.0) = 1.5% and μ_1/σ_1 = 5.0/25.0 = 20.0%. The correlation coefficient across all four securities (calculated in Excel) is TC = 0.58. The transfer coefficient for the optimal active weights is by definition 1.0 but can be verified by the calculated correlation coefficient. The risk-weighted values for Security #1 are then $\Delta w_1 \sigma_1$ = 0.18(25.0) = 4.5% and μ_1/σ_1 = 5.0/25.0 = 20.0%.

2. The forecasted active return of the optimal portfolio is the sum of the active weights times active returns for each security: 0.18(5.0) + 0.09(10.0) + (–0.18)(–5.0) + (–0.09)(–10.0) = 3.6%. The active risk of the optimal portfolio is the square root of the sum of active weights squared times the active volatility squared for each security: $[(0.18)^2(25.0)^2 + (0.09)^2(50.0)^2 + (-0.18)^2(25.0)^2 + (-0.09)^2(50.0)^2]^{1/2}$ = 9.0%. Calculate the forecasted active return and active risk of the managed portfolio using the *actual* rather than unconstrained optimal active weights.

Solution:

The forecasted active return of the managed portfolio is 0.06(5.0) + 0.04(10.0) + 0.07(–5.0) + (–0.17)(–10.0) = 2.1%. The active risk of the managed portfolio is the square root of the sum of actual active weights squared times the active volatility squared for each security, $[(0.06)^2(25.0)^2 + (0.04)^2(50.0)^2 + (0.07)^2(25.0)^2 + (-0.17)^2(50.0)^2]^{1/2}$ = 9.0%, as specified by the active risk constraint.

3. Verify the expanded fundamental law of active management using the active portfolio return, active portfolio risk, and transfer coefficient calculations in Parts 1 and 2. The individual active return forecasts and optimal active weights were sized using an information coefficient of IC = 0.20 and breadth of BR = 4.

Solution:

The expanded fundamental law states that the expected portfolio active return will be $E(R_A) = (TC)(IC)\sqrt{BR}\,\sigma_A = 0.58 \times 0.20 \times (4)^{1/2} \times 9.0 = 2.1\%$, consistent with the direct calculation in the Solution to 2.

We close this sub-section by noting that the transfer coefficient, TC, also comes into play when calculating the optimal amount of active risk for an actively managed portfolio with constraints. Specifically, with constraints and using notation consistent with expressions in the fundamental law, Equation 8 becomes

$$\sigma_A = TC\frac{IR^*}{SR_B}\sigma_B,$$

where IR* is the information ratio of an otherwise unconstrained portfolio. Employing this optimal level of aggressiveness leads to a maximum possible value of the constrained portfolio's squared Sharpe ratio:

$$SR_P^2 = SR_B^2 + (TC)^2(IR^*)^2.$$

As noted previously, the active risk of an actively managed fund can be adjusted to its optimal level while preserving the information ratio by adding long or short positions in the benchmark portfolio. For further insight, note that with a transfer coefficient of 0.00, the optimal amount of active risk calculated is zero. In other words, the investor should just invest in the benchmark portfolio.

We now illustrate the impact of the transfer coefficient with the following example. If the actively managed portfolio has a transfer coefficient of 0.50 and an unconstrained information ratio of 0.30 and the benchmark portfolio has a Sharpe ratio of 0.40 and risk of 16.0%, then the optimal amount of aggressiveness in the actively managed portfolio is 0.50(0.30/0.40)16.0 = 6.0%. If the actively managed portfolio is constructed with this amount of active risk, the Sharpe ratio will be $(0.40^2 + 0.50^2 \times 0.30^2)^{1/2} = 0.43$. If the constrained portfolio has an active risk of 8.0%, the active risk can be lowered to the optimal level of 6.0% by mixing 1 − 6.0/8.0 = 25% in the benchmark and 75.0% in the actively managed fund.

Ex Post Performance Measurement

Most of the fundamental law perspectives discussed up to this point relate to the expected value added through active portfolio management. Actual performance in any given period will vary from its expected value in a range determined by the benchmark tracking risk. We now turn our attention to examining actual performance, the *ex post* analysis of the realized value added.

The key determinant of the sign and magnitude of the realized value added in Equation 3 is the degree to which the portfolio has positive active weights on securities that realize positive relative returns and negative active weights on securities that realize negative relative returns. In other words, actual performance is measured by the relationship between relative weights and realized relative returns. Knowing how actual returns match up with realized returns (the *realized* information coefficient, IC_R) allows the investor to examine what realized return to expect given the transfer coefficient. Specifically, expected value added conditional on the realized information coefficient, IC_R, is

$$E\left(R_A \middle| IC_R\right) = (TC)\left(IC_R\right)\sqrt{BR}\,\sigma_A. \tag{15}$$

Equation 15 is similar to the fundamental law, shown in Equation 14, but in Equation 15 the realized information coefficient, IC_R, replaces the *expected* information coefficient, IC.

We can represent any difference between the actual active return of the portfolio and the conditional expected active return with a noise term:

$$R_A = E(R_A \mid IC_R) + \text{Noise}. \tag{16}$$

Equation 16 states that the realized value added of an actively managed portfolio can be divided into two parts. The first part comes from the expected value added given the realized skill of the investor that period. The second part represents any noise that results from constraints that impinge on the optimal portfolio structure.

Equation 15 also leads to an *ex post* (i.e., realized) decomposition of the portfolio's active return variance into two parts: variation due to the realized information coefficient and variation due to constraint-induced noise. Clarke, de Silva, and Thorley (2005) showed that the two parts of the realized variance are proportional to TC^2 and $1 - TC^2$. For example, with a TC value of, say, 0.60, only TC^2 = 36% of the realized variation in performance is attributed to variation in the realized information coefficient, and $1 - TC^2$ = 64% comes from constraint-induced noise. Low-TC investors will frequently experience periods when the forecasting process succeeds but actual performance is poor or when actual performance is good even though the return-forecasting process fails.

EXAMPLE 7

Ex Post Performance

Consider an active management strategy that includes BR = 100 investment decisions (e.g., 100 individual stocks, whose active returns are uncorrelated, and annual rebalancing), an expected information coefficient of IC = 0.05, a transfer coefficient of TC = 0.80, and annualized active risk of σ_A = 4.0%. Thus, the expected value added according to the fundamental law is

$$E\left(R_A\right) = (TC)(IC)\sqrt{BR}\,\sigma_A = 0.80 \times 0.05 \times \sqrt{100} \times 4.0\% = 1.6\%.$$

1. Suppose that the *realized* information coefficient in a given period is −0.10, instead of the expected value of IC = 0.05. In the absence of constraint-induced noise, what would be the value added that period?

Solution:

The value added, without including constraint-induced noise (which has an expected value of zero) is

$$E\left(R_A \middle| IC_R\right) = (TC)\left(IC_R\right)\sqrt{BR}\,\sigma_A = 0.80 \times (-0.10) \times \sqrt{100} \times 4.0\%$$
$$= -3.2\%.$$

In other words, conditional on the actual information coefficient, the investor should expect an active return that is negative because the realized information coefficient is negative.

2. Suppose that the actual return on the active portfolio was −2.6%. Given the −0.10 realized information coefficient, how much of the forecasted active return was offset by the noise component?

Solution:

The noise portion of the active return is the difference between the actual active return and the forecasted active return: −2.6 − (−3.2) = 0.6%. In other words, the noise component helped offset the negative value added from poor return forecasting. Of course, the constraint-induced noise component could just as easily have gone the other way, exacerbating the negative value added. Note that the negative realized active return of −2.6% is well within the range associated with the tracking error (active risk) of 4.0% per period.

3. What percentage of the performance variance (i.e., tracking risk squared) in this strategy over time is attributed to variation in the realized information coefficient (i.e., forecasting success), and what percentage of performance variance is attributed to constraint-induced noise?

Solution:

Given the transfer coefficient of TC = 0.80, TC^2 = 64% of the variation in performance over time is attributed to the success of the forecasting process, leaving 36% due to constraint-induced noise.

APPLICATIONS OF THE FUNDAMENTAL LAW AND GLOBAL EQUITY STRATEGY

7

- ☐ explain how the information ratio may be useful in investment manager selection and choosing the level of active portfolio risk
- ☐ compare active management strategies, including market timing and security selection, and evaluate strategy changes in terms of the fundamental law of active management

In this section, we discuss three specific applications of active portfolio management: one application to a global equity strategy with different sets of active return forecasts and constraints and two applications to US fixed income. These applications will further illustrate how the fundamental law is used to evaluate active portfolio strategies, including security selection and market timing.

Global Equity Strategy

In our first example, we show how the fundamental law can be used to calculate the expected active return for an actively managed portfolio benchmarked to the MSCI All Country World Index (ACWI). This global equity example focuses on the cross-sectional characteristics of the fundamental law, whereas the US fixed-income examples that follow also include time-series implications of the law. The investable assets in this example are the individual MSCI market indexes—including the 21 EAFE (Europe, Australasia, and Far East) markets, the United States, Canada, and the Emerging Markets Index—for a total of 24 assets. "Now" is the beginning of the calendar year 2019. For purposes of illustration, we will assume that the future will be like the past in terms of active risk, and thus we will base our estimates on the US dollar return to the MSCI market indexes from 2009 to 2018. In practice, managerial judgment or a commercial model would be used to forecast risk. The various rankings of the markets' forecasted active returns for the calendar year 2019 are hypothetical.

The *ex ante* expected active risk of each asset is equal to the annualized historical standard deviation of beta-adjusted differences between the individual market return and the ACWI return, as shown in the third column of Exhibit 7. For example, the active risk of the United Kingdom is 6.4% and the active risk of Japan is 9.1%. Note that the risk estimates are for active returns (i.e., the difference between the individual asset and benchmark returns). The total risk of each market would be higher based on the estimated risk of the benchmark and the benchmark beta.

Exhibit 7: Long–Short Global Equity Fund for 2019 (risk statistics based on MSCI returns from 2009 to 2018)

Market	Score	Active Return Volatility	Expected Active Return	Active Weight	ACWI Benchmark Weight	Portfolio Weight
United Kingdom	2.0	6.4%	1.3%	15.7%	5.2%	20.9%
Japan	0.0	9.1%	0.0%	−2.0%	7.6%	5.7%
France	−2.0	8.5%	−1.7%	−12.0%	3.4%	−8.5%
Germany	0.0	8.4%	0.0%	2.8%	2.7%	5.6%
Switzerland	2.0	7.8%	1.6%	9.7%	2.7%	12.3%

Market	Score	Active Return Volatility	Expected Active Return	Active Weight	ACWI Benchmark Weight	Portfolio Weight
Australia	0.0	11.1%	0.0%	−1.0%	2.1%	1.1%
Spain	−2.0	16.0%	−3.2%	−5.3%	1.0%	−4.3%
Sweden	0.0	10.1%	0.0%	−1.2%	0.8%	−0.4%
Hong Kong SAR	1.0	12.0%	1.2%	2.3%	1.2%	3.5%
Netherlands	0.0	7.9%	0.0%	1.8%	1.1%	2.9%
Italy	−1.0	15.1%	−1.5%	−0.8%	0.7%	−0.1%
Singapore	0.0	11.8%	0.0%	−1.8%	0.4%	−1.3%
Belgium	1.0	10.2%	1.0%	4.4%	0.3%	4.7%
Denmark	0.0	12.3%	0.0%	−1.2%	0.5%	−0.7%
Finland	−1.0	13.7%	−1.4%	−2.5%	0.3%	−2.2%
Norway	0.0	13.7%	0.0%	0.1%	0.2%	0.3%
Israel	1.0	15.3%	1.5%	1.4%	0.2%	1.6%
Ireland	0.0	14.9%	0.0%	0.3%	0.2%	0.5%
Austria	−1.0	14.6%	−1.5%	−1.6%	0.1%	−1.6%
Portugal	0.0	15.2%	0.0%	0.9%	0.0%	0.9%
New Zealand	1.0	14.2%	1.4%	1.8%	0.1%	1.8%
United States	0.0	3.8%	0.0%	−5.3%	54.3%	48.9%
Canada	−1.0	9.2%	−0.9%	−6.7%	3.0%	−3.7%
Emerging	0.0	9.0%	0.0%	0.1%	11.9%	12.0%
Total	0.0			0.0%	100.0%	100.0%

	Transfer Coefficient	Information Coefficient	Breadth
	0.995	0.099	24.5

	Active Return	Active Risk	Information Ratio
	0.98%	2.00%	0.49

The 24 individual assets in Exhibit 7 are listed approximately by size in the EAFE benchmark, followed by the United States, Canada, and the emerging markets. For example, the United Kingdom has a benchmark weight of 5.2% and Canada has a benchmark weight of 3.0%. Scores representing an active investor's forecasts of the relative performance of each asset during 2019 are assigned to each market. The scores are one of five numerical values that represent a managerial forecast of strong outperformance (2.0), weak outperformance (1.0), neutral performance (0.0), weak underperformance (−1.0), and strong underperformance (−2.0). The number of scores in each of these five categories is based on the requirement that the scores sum to zero and have a cross-sectional standard deviation of 1.

The active return forecasts in the fourth column of Exhibit 7 are based on the Grinold rule in Equation 10 of "IC times volatility times score," where IC is the *ex ante* information coefficient that measures the assumed accuracy of the investor's relative rankings, as illustrated by the right leg of the correlation triangle in Exhibit 6. In this example, we use an assumed information coefficient of 0.10; thus, forecasted and realized active security returns are expected to have a cross-sectional correlation coefficient of 0.10. For example, the active return forecast for the United Kingdom, which has a score of 2.0, is 0.10(6.4)(2.0) = 1.3%. Alternatively, the active return forecast for Japan is 0.0% because the score is 0. As explained later, the information coefficient used in fundamental law accounting will be adjusted down to 0.095, as shown at the bottom of Exhibit 7, to account for the assignment of scores in this particular example.

The active weights for each market are based on the active return forecast and a numerical optimizer (i.e., Excel Solver) with the objective to maximize the expected active return of the portfolio, subject to a 2.00% constraint on active risk. Note that while the active weights for each market are generally correlated with the forecasted active returns in Exhibit 7, they are not perfectly proportional for two reasons. First, the optimizer also takes into account the estimated correlations between each market's active return, based on the MSCI monthly return data from 2009 to 2018. Exhibit 8 reports the estimated active return correlations for the eight largest EAFE countries; the full correlation matrix is not reported to conserve space. For example, the correlation coefficient between the United Kingdom (GB) and Japan (JP) is fairly low at −0.02, while the correlation coefficient between France (FR) and Germany (DE) is higher at 0.30. Note that these correlation coefficients are for *active* returns (i.e., the differences between the individual market and ACWI benchmark returns). The correlations for *total* market returns would all be positive and much higher—for example, values that range from 0.4 to 0.9.

Exhibit 8: Active Return Correlation Coefficients for Eight Countries (based on MSCI returns from 2009 to 2018)

Country	GB	JP	FR	DE	CH	AU	ES	SE	
GB	1.000	−0.02	0.21	0.08	0.13	0.00	0.18	0.14	…
JP	−0.02	1.000	−0.08	−0.03	0.04	−0.08	0.01	−0.07	…
FR	0.21	−0.08	1.000	0.30	0.16	−0.03	0.34	0.15	…
DE	0.08	−0.03	0.30	1.000	0.10	−0.07	0.19	0.15	…
CH	0.13	0.04	0.16	0.10	1.000	0.06	0.11	0.10	…
AU	0.00	−0.08	−0.03	−0.07	0.06	1.000	0.01	0.06	…
ES	0.18	0.01	0.34	0.19	0.11	0.01	1.000	0.11	…
SE	0.14	−0.07	0.15	0.15	0.10	0.06	0.11	1.000	…
	…	…	…	…	…	…	…	…	…

The second reason that the active weights in Exhibit 7 are not perfectly proportional to the forecasted active returns is that the active weights are constrained by the optimizer to sum to zero. For example, the highest active weight in Exhibit 7 is for the United Kingdom, at 15.7%, and the lowest active weight is for France, at −12.0%. These active weights are added to the benchmark weights to give the total portfolio weights in the last column of Exhibit 7. For example, the total weight for the United States is 48.9%, even though the active weight is −5.3%, because the US benchmark weight in the ACWI is 54.3%. In fact, the optimization in Exhibit 7 is for a relatively unconstrained long–short portfolio where the sum of the positive total weights is about 120% and the sum of the negative total weights is about −20%, what might be called a "120/20 long–short" strategy in practice.

Because the optimization is basically unconstrained, the transfer coefficient or risk-weighted correlation between active return forecasts and active weights shown at the bottom of Exhibit 7 is 0.995, almost perfect. The transfer coefficient in this example takes into account all the risk statistics (i.e., forecasted active volatilities *and* forecasted correlations), but it is not exactly 1.0 because of the budget constraint that the active weights sum to zero. Alternatively, if the sum of active weights were allowed to be non-zero, effectively allowing for risk-free cash or leverage in the equity portfolio to meet the budget constraint, the transfer coefficient would be exactly 1.0. The breadth of the strategy shown at the bottom of Exhibit 7 is 24.5, slightly higher than the number of individual assets, 24.0, because the risk model includes active return correlation coefficients that are different from zero. If all the off-diagonal correlations in the extended table in Exhibit 8 were exactly zero, then breadth would be exactly 24.0, instead of 24.5.

The fundamental law in Equation 14 states that the expected active return on the portfolio is $E(R_A) = (TC)(IC)\sqrt{BR}\,\sigma_A = 0.995 \times 0.099 \times (24.5)^{1/2} \times 2.00 = 0.98\%$. Alternatively, the expected active return of 0.98%, shown at the bottom of Exhibit 7, is calculated as the sum of the active weights times active returns. Thus, the accuracy of the fundamental law is quite high. The fundamental law is often expressed in terms of the information ratio, or forecasted active return over active risk. Using this framework, the validation of the fundamental law is $IR = (TC)(IC)\sqrt{BR} = 0.995 \times 0.099 \times (24.5)^{1/2} = 0.49$, equal to actual forecasted active return divided by active risk, 0.98/2.00 = 0.490. Because the information ratio in this relatively unconstrained portfolio is unaffected by the aggressiveness of the strategy, we would get the same IR value if the active risk were allowed to be higher. For example, if the active risk specified to the optimizer were increased to 3.00%, the forecasted active return would increase to 1.47%, an information ratio of, again, 1.47/3.00 = 0.49.

Exhibit 9 continues examining the global equity strategy but uses a slightly different assignment of scores than Exhibit 7 to illustrate how this change affects the values in the fundamental law. Specifically, the scores for Germany (DE) and the United Kingdom (UK) have been switched as well as the scores for Switzerland (CH) and Australia (AU). While the breadth in Exhibit 9 is unchanged at 24.5, the information coefficient has increased slightly to 0.105, compared with 0.099 in Exhibit 7. Even though the assumed IC used to create the expected active returns in Exhibit 9 is still 0.10, the IC used in fundamental law accounting has increased because the new assignment of scores represents a slightly more ambitious forecast. For example, the active (i.e., benchmark relative) returns for France and Germany in Exhibit 9 are now forecasted to go strongly in opposite directions, even though they are positively correlated according to the risk model in Exhibit 8. Given the increase in IC and slight change in TC, the fundamental law calculation for Exhibit 9 is now $IR = (TC)(IC)\sqrt{BR} = 0.997 \times 0.105 \times (24.5)^{1/2} = 0.532$, equal to the actual value of 0.52.

Exhibit 9: Long–Short Global Equity Fund with Different Scores for 2019 (risk statistics based on MSCI returns from 2009 to 2018)

Market	Score	Active Return Volatility	Expected Active Return	Active Weight	ACWI Benchmark Weight	Portfolio Weight
United Kingdom	0.0	6.4%	0.0%	1.7%	5.2%	6.9%
Japan	0.0	9.1%	0.0%	−0.3%	7.6%	7.3%
France	−2.0	8.5%	−1.7%	−11.4%	3.4%	−8.0%
Germany	2.0	8.4%	1.7%	14.5%	2.7%	17.2%
Switzerland	0.0	7.8%	0.0%	−2.0%	2.7%	0.6%
Australia	2.0	11.1%	2.2%	7.7%	2.1%	9.9%
Spain	−2.0	16.0%	−3.2%	−4.9%	1.0%	−4.0%
Sweden	0.0	10.1%	0.0%	−1.2%	0.8%	−0.4%
Hong Kong SAR	1.0	12.0%	1.2%	2.8%	1.2%	4.0%
Netherlands	0.0	7.9%	0.0%	3.0%	1.1%	4.1%
Italy	−1.0	15.1%	−1.5%	−0.9%	0.7%	−0.2%

Market	Score	Active Return Volatility	Expected Active Return	Active Weight	ACWI Benchmark Weight	Portfolio Weight
Singapore	0.0	11.8%	0.0%	−1.4%	0.4%	−1.0%
Belgium	1.0	10.2%	1.0%	5.8%	0.3%	6.1%
Denmark	0.0	12.3%	0.0%	−0.7%	0.5%	−0.1%
Finland	−1.0	13.7%	−1.4%	−3.2%	0.3%	−2.9%
Norway	0.0	13.7%	0.0%	1.3%	0.2%	1.5%
Israel	1.0	15.3%	1.5%	1.5%	0.2%	1.7%
Ireland	0.0	14.9%	0.0%	0.2%	0.2%	0.4%
Austria	−1.0	14.6%	−1.5%	−1.8%	0.1%	−1.7%
Portugal	0.0	15.2%	0.0%	1.4%	0.0%	1.4%
New Zealand	1.0	14.2%	1.4%	1.4%	0.1%	1.4%
United States	0.0	3.8%	0.0%	−5.0%	54.3%	49.2%
Canada	−1.0	9.2%	−0.9%	−5.3%	3.0%	−2.3%
Emerging	0.0	9.0%	0.0%	−2.9%	11.9%	9.0%
Total	0.0			0.0%	100.0%	100.0%

	Transfer Coefficient	Information Coefficient	Breadth
	0.997	0.105	24.5

	Active Return	Active Risk	Information Ratio
	1.04%	2.00%	0.52

We now apply constraints to the global equity strategy to focus on the transfer coefficient. Specifically, Exhibit 10 shows two *constrained* portfolio optimizations using the same score assignments and thus active return forecasts as in Exhibit 7. The first optimization, shown on the left-hand side of Exhibit 10, has two constraints. First, the portfolio is constrained to be long only (i.e., a negative active weight for any given market cannot be bigger than the benchmark weight). For example, France has an active weight of −3.4%, bounded by the benchmark weight of 3.4%, so that the total weight for France in the managed portfolio is zero. Second, the portfolio weights are constrained to not be more than 10.0% over or under the benchmark weight (i.e., the absolute value of any given market active weight cannot be greater than 10.0%). For example, the active weights for the United Kingdom and Switzerland are limited to 10.0% and the active weight for the United States is limited to −10.0%.

Exhibit 10: Constrained Global Equity Funds for 2019 (risk statistics based on MSCI returns from 2009 to 2018)

Market	Active Weight	ACWI Benchmark Weight	Portfolio Weight	Active Weight	ACWI Benchmark Weight	Portfolio Weight
United Kingdom	10.0%	5.2%	15.2%	8.6%	5.2%	15.2%
Japan	−6.6%	7.6%	1.0%	−7.8%	7.76%	0.0%
France	−3.4%	3.4%	0.0%	−3.7%	3.4%	0.0%
Germany	−2.7%	2.7%	0.0%	−3.5%	2.7%	0.0%
Switzerland	10.0%	2.7%	12.7%	10.0%	2.7%	12.7%
Australia	−2.1%	2.1%	0.0%	−2.8%	2.1%	0.0%

Market	Active Weight	ACWI Benchmark Weight	Portfolio Weight	Active Weight	ACWI Benchmark Weight	Portfolio Weight
Spain	−1.0%	1.0%	0.0%	−1.2%	1.0%	0.0%
Sweden	−0.8%	0.8%	0.0%	−1.2%	0.8%	0.0%
Hong Kong SAR	5.9%	1.2%	7.1%	9.5%	1.2%	5.7%
Netherlands	−1.1%	1.1%	0.0%	−1.0%	1.1%	0.0%
Italy	−0.7%	0.7%	0.0%	−0.8%	0.7%	0.0%
Singapore	−0.4%	0.4%	0.0%	−0.5%	0.4%	0.0%
Belgium	6.1%	0.3%	6.4%	−0.4%	0.3%	0.0%
Denmark	−0.5%	0.5%	0.0%	−0.4%	0.5%	0.0%
Finland	−0.3%	0.3%	0.0%	−0.3%	0.3%	0.0%
Norway	−0.2%	0.2%	0.0%	−0.3%	0.2%	0.0%
Israel	4.2%	0.2%	4.4%	10.0%	0.2%	10.2%
Ireland	−0.2%	0.2%	0.0%	−0.1%	0.2%	0.0%
Austria	−0.1%	0.1%	0.0%	−0.1%	0.1%	0.0%
Portugal	0.0%	0.0%	0.0%	−0.1%	0.0%	0.0%
New Zealand	4.7%	0.1%	4.8%	10.0%	0.1%	10.1%
United States	−10.0%	54.3%	44.3%	−10.0%	54.3%	44.3%
Canada	−3.0%	3.0%	0.0%	−3.7%	3.0%	0.0%
Emerging	−7.7%	11.9%	4.2%	−10.0%	11.9%	1.9%
Total	0.0%	100.0%	100.0%	0.0%	100.0%	100.0%

	Transfer Coefficient	Information Coefficient	Breadth	Transfer Coefficient	Information Coefficient	Breadth
	0.694	0.099	24.5	0.567	0.099	24.5

	Active Return	Active Risk	Information Ratio	Active Return	Active Risk	Information Ratio
	0.68%	2.00%	0.34	0.76%	2.74%	0.28

The long-only and maximum over- or underweight constraints substantially reduce the transfer of active return forecasts into active weights, as shown by the transfer coefficient of 0.694 at the bottom of the left side of Exhibit 10, compared with 0.995 for the same scores and active return forecasts in Exhibit 7. The impact of this transfer coefficient on expected active return according to the fundamental law is $E(R_A) = (TC)(IC)\sqrt{BR}\,\sigma_A$ = 0.694 × 0.099 × $(24.5)^{1/2}$ × 2.00 = 0.68%, compared with 0.98% for the unconstrained portfolio in Exhibit 7. Similarly, the impact of this transfer coefficient measured by the information ratio is $IR = (TC)(IC)\sqrt{BR}$ = 0.694 × 0.099 × $(24.5)^{1/2}$ = 0.34, compared with 0.49 for the unconstrained portfolio. In other words, the expected active return and information ratio are reduced by almost a third because of the constraints imposed in portfolio construction.

As previously mentioned, an increase in the allowed active risk from 2.00% to 3.00% in the unconstrained portfolio in Exhibit 7 proportionally increases the active return, leaving the information ratio at about 0.49. However, an increase in allowed active risk to 3.00% does *not* preserve the information ratio of the constrained portfolio, as shown by the optimization on the right-hand side of Exhibit 10. Specifically, the higher active risk leads to more variation in unconstrained active weights, as shown in Equation 11; thus, the constraints become more binding. For example, the active weight for New Zealand, which is 4.7% on the left-hand side of Exhibit 10, is capped at the maximum possible value of 10.0% on the right-hand side of Exhibit 10. The result is a further reduction in the transfer coefficient from 0.694 to 0.567, leading to a reduction in the information ratio to IR = $(TC)\ (IC)\sqrt{BR} = 0.567 \times 0.099 \times (24.5)^{1/2}$ = 0.28, compared with 0.34 at the lower active portfolio risk of 2.0%.

The key concept is that although an unconstrained IR is invariant to the level of active risk, as shown by the dark line in Exhibit 5, the IR for a *constrained* portfolio generally decreases with the aggressiveness of the strategy. Specifically, the dark line in Exhibit 5 for a constrained portfolio would curve downward from left to right in accordance with an increasingly lower transfer coefficient. Thus, the constraints that are imposed on the portfolio should inform the decision of how aggressively to apply an active management strategy.

EXAMPLE 8

Compare and Contrast Active Management Strategies

Consider two active management strategies: individual stock selection, with a benchmark composed of 100 securities, and industrial sector selection, with a benchmark of nine sectors. The active security returns are defined as residuals in a risk model and thus are essentially uncorrelated, and forecasts are independent from year to year. Suppose the individual stock investor is expected to exhibit skill as measured by an information coefficient of 0.05, while the industrial sector investor has a higher information coefficient of 0.15.

1. Conceptually, what is the breadth (i.e., number of independent decisions per year) of each active management strategy?

Solution:

Given that the active asset returns in each strategy are uncorrelated and forecasts are independent from year to year, the breadth of the security selection strategy is BR = 100 and of the sector selection strategy is BR = 9.

2. Calculate the expected information ratio for each strategy under the assumption that each investor's forecasts can be implemented without constraints, such as the long-only constraint or a limit on turnover each year.

Solution:

The expected information ratio of the unconstrained security selection strategy is calculated as IR = $(IC)\sqrt{BR} = 0.05 \times \sqrt{100} = 0.50$, while the information ratio of the industrial sector selection strategy is IR = $(IC)\sqrt{BR} = 0.15 \times \sqrt{9} = 0.45$.

3. Suppose the aggressiveness of each active management strategy is established by a portfolio active risk target of 3.0% per year. What is the expected active return to each strategy?

Solution:

The expected active return to the unconstrained security selection strategy is 0.50(3.0) = 1.50%, while the expected active return of the industrial sector selection strategy is 0.45(3.0) = 1.35%.

4. Under the more realistic assumption that the individual security selection strategy is constrained to be long only and has turnover limits, the transfer

coefficient has a value of 0.60. Calculate the constrained information ratio and expected active return of the security selection strategy.

Solution:

The information ratio of the constrained security selection strategy is $IR = (TC)(IC)\sqrt{BR} = 0.60 \times 0.05 \times \sqrt{100} = 0.30$, rather than 0.50, and the expected active return is 0.30(3.0) = 0.90%, rather than 1.50%.

5. Suppose the aggressiveness of the constrained individual security selection strategy is increased to a portfolio active risk target of 4.0% per year. Conceptually, what is likely to happen to the information ratio, and why?

Solution:

A more aggressive implementation of the constrained security selection strategy will likely result in larger deviations of constrained weights from unconstrained weights and thus a lower transfer coefficient. For example, the transfer coefficient might drop from 0.60 to 0.50, leading to an information ratio of only $IR = (TC)(IC)\sqrt{BR} = 0.50 \times 0.05 \times \sqrt{100} = 0.25$. Thus, instead of a proportional increase in the expected active return associated with an increase in the active portfolio risk from 3.0% to 4.0%, the expected active return would only increase from 0.9% to 0.25(4.0) = 1.0%.

8 FIXED-INCOME STRATEGIES

- ☐ explain how the information ratio may be useful in investment manager selection and choosing the level of active portfolio risk
- ☐ compare active management strategies, including market timing and security selection, and evaluate strategy changes in terms of the fundamental law of active management

Two additional examples of the fundamental law in practice are based on the Bloomberg Barclays US fixed-income index returns. Consider first an active management strategy of over- and underweighting credit exposure once a quarter using corporate investment-grade and high-yield bond portfolios as assets. Let the benchmark portfolio be composed of 70% investment-grade bonds and 30% high-yield bonds. Each quarter, the active investor makes a single dichotomous decision either to overweight the investment-grade asset (and thus underweight the high-yield asset) or to overweight the high-yield asset (and thus underweight the investment-grade asset). In addition to switching to a fixed-income example, we are also now moving into a time-series application of the fundamental law instead of the purely cross-sectional application.

For example, consider two bond portfolios, an investment-grade portfolio and a high-yield portfolio. The quarterly return volatility of the IG (investment-grade) asset is 2.84%, and the quarterly return volatility of the HY (high-yield) asset is 4.64%, with an estimated correlation between the two of 0.575. The *active* risk of this decision is the volatility of the differential returns between the two bond portfolios, $[(2.84)^2 - 2(2.84)(4.64)(0.575) + (4.64)^2]^{1/2} = 3.80\%$. In effect, the active investor assigns a "score" of either +1.0 or −1.0 on credit exposure each quarter, with an *annualized* active risk of $3.80 \times (4)^{1/2} = 7.60\%$. Suppose the fixed-income investor expects to call the market correctly 55% of the time (i.e., 11 out of 20 quarters). If the investor makes the correct decision 55% of the time and an incorrect decision 45% of the time, then the time-series information coefficient is 0.55 − 0.45 = 0.10.

If a time series of T predicted dichotomous (i.e., plus or minus 1.0) scores, $S_{P,t}$, and a time series of T realized dichotomous scores, $S_{R,t}$, both have zero means, then the time-series covariance between the two is $\text{COV}(S_P, S_R) = \frac{1}{T}\sum_{t=1}^{T} S_{P,t} S_{R,t}$.

The product of the two scores at time period t is 1.0 if the scores have the same sign (i.e., the decision is correct) and −1.0 if the scores have different signs (i.e., the decision is incorrect). Because the scores have unit variances, the correlation coefficient is equal to the covariance. Thus, the time-series correlation is equal to the number of correct decisions minus the number of incorrect decisions all over total decisions, or, in other words, the percentage correct minus the percentage incorrect.

Without a limit on active risk, the expected active return can be calculated using a simple probability-weighted average: 0.55(3.80) + 0.45(−3.80) = 38 bps per quarter. But to illustrate the fundamental law, we use the Grinold rule in Equation 10 of "alpha equals IC times volatility times score": 0.10(3.80)(1.0) = 38 bps.

The investor decides to limit the annual active risk to 2.00% and thus sets the active weight (i.e., deviation from the 70/30 benchmark weights) at 2.00/7.60 = 26.3%. Under the assumption that active returns are uncorrelated over time, the breadth of this strategy is 4.0, the four quarterly rebalancing decisions made each year. Thus, in quarters when the investor believes credit risk will pay off, the managed portfolio is invested 70.0% − 26.3% = 43.7% in investment-grade bonds and 30% + 26.3% = 56.3% in high-yield bonds. Alternatively, in quarters where the investor believes credit risk will not pay off, the active portfolio has 70.0% + 26.3% = 96.3% in investment-grade bonds and only 30% − 26.3% = 3.7% in high-yield bonds. According to the simple form of the fundamental law, the expected annualized active return to this strategy is $E(R_A) = (\text{IC})\sqrt{\text{BR}}\,\sigma_A = 0.10 \times (4.0)^{1/2} \times 2.00$ = 40 bps a year, or 10 bps per quarter. Alternatively, given the active weight of 26.3% motivated by the desire to limit active risk, the expected quarterly return can be calculated more directly as 0.263 × 38 = 10 bps. Given the small breadth of this strategy, the annual information ratio is only $\text{IR} = (\text{IC})\sqrt{\text{BR}} = 0.10 \times (4.0)^{1/2} = 0.20$.

The key concept in this illustration is that the breadth of the strategy is only 4, meaning four active management decisions per year. The same small-breadth problem also applies to quarterly tactical asset allocation decisions in a simple strategy that switches between equity and cash. There are so few opportunities to make an active decision in these "market-timing" strategies that the investor's accuracy as measured by the information coefficient must be quite high to achieve even a modest information ratio. A full description of the breadth calculation requires relatively complex matrix formulas that take into account the correlations between security returns. However, one "rule of thumb" is that breadth is approximately $\text{BR} = N/[1 + (N - 1)\rho]$, where N is the number of securities and ρ is the average correlation between the active security returns. In this fixed-income example, ρ = 0.0, so breadth is BR = 4.0.

A natural question is whether the expected information ratio can be increased by switching more frequently—say, monthly. Although it is somewhat more complicated to show, the basic answer is yes—*if* the information coefficient of 0.10 can be maintained and *if* the credit exposure decisions in this example are truly independent over time. For example, making monthly decisions that do not change during the quarter (i.e., signals of +1.0, +1.0, and +1.0 in January, February, and March) will *not* increase the information ratio of 0.20. However, if the monthly signals are truly uncorrelated with each other, then the information ratio in this example would be IR = $(IC)\sqrt{BR} = 0.10 \times \sqrt{12} = 0.35$. Although somewhat implausible, if an investor made daily decisions (250 trading days a year) that were truly independent and were *still* correct 55% of the time, the expected information ratio could potentially increase to IR = $(IC)\sqrt{BR} = 0.10 \times \sqrt{250} = 1.58$.

The high 1.58 information ratio indicates that the investor could earn an expected active return of 3.16% with active risk of only 2.00%. With such a high information ratio, the investor might be inclined to increase the aggressiveness of the credit risk strategy—for example, doubling to an expected return of 2 × 3.16% = 6.32% and active risk of 2 × 2.00% = 4.00%. Besides the issue of transaction costs, this more aggressive strategy would likely bump up against various constraints. For example, at the higher 4.00% active risk, the required active weights would be plus and minus 4.00/7.60 = 52.6%. In other words, a tilt against credit risk would require a total portfolio weight of 70% + 52.6% = 122.6% in investment-grade bonds funded by a −22.6% *short* position in high-yield bonds.

The essential logic of this example is not confined to a dichotomous decision; the same general perspectives would hold if the single credit risk signal were continuous—for example, numbers like −0.57 or 1.32. Then under the more aggressively applied active risk target of 4.0%, a signal of −0.57 would require an active weight of −0.57(4.0)/7.6 = −30.0%. With a benchmark portfolio of 70.0% investment-grade and 30.0% high-yield bonds, this active weight translates into a 100% position in investment-grade bonds and no position in high-yield bonds. Alternatively, for a positive credit risk signal of 1.32, the required active weight would be 1.32(4.00)/7.60 = 69.5% (i.e., 100% in high-yield bonds and almost no position in investment-grade bonds). In other words, for this more aggressive strategy under a long-only constraint, the transfer coefficient would be less than 1 and the expanded fundamental law, IR = $(TC)(IC)\sqrt{BR}$, would come into play. Under a normal distribution for scores, the transfer coefficient of this strategy is 0.62, so the expected information ratio is only IR = $0.62 \times 0.10 \times \sqrt{250} = 0.98$, not 1.58. For an active risk of 4.00%, the expected active return is thus only 0.98 × 4.00% = 3.92%, not 6.32%. Please note that the transfer coefficient in this example is based on the calculation $\Phi(1.32) - \Phi(-0.57) = 0.62$, where $\Phi(S)$ is the cumulative standard normal distribution function. Given long-only limits on positions, the actual active risk of the constrained portfolio would be lower than 4.0%. In other words, the actual active weights (determined by a numerical optimizer) would need to be larger than the simple formula (S)4.0/7.6 to get back up to an actual active risk of 4.0%.

For our second fixed-income example, consider an active management strategy using the five US Treasury bond portfolios in Exhibit 11 as the individual assets. Let the neutral benchmark be an equally weighted composite portfolio of the five, or 20% invested in each asset, but with annual rebalancing. In other words, we are now moving back into a purely cross-sectional application of the fundamental law.

Exhibit 11: Bloomberg Barclays US Treasury Bond Average Returns and Risk (return statistics from 2009 to 2018)

	Treas. 0–1	Treas. 1–3	Treas. 3–7	Treas. 7–10	Treas. 10–20
Avg. Ret.	0.40%	0.90%	2.21%	3.15%	3.89%
Volatility	0.17%	0.85%	3.20%	5.86%	7.95%

Exhibit 12 shows the volatility of the historical return differences between each asset and the equally weighted benchmark. Note that while the absolute volatility of each asset return goes up with maturity in Exhibit 11, the *active* volatility with respect to the benchmark is highest for the assets with the shortest maturity, at 3.45%, and the longest maturity, at 4.57%. Exhibit 12 also shows the estimated active (i.e., benchmark relative) return correlation matrix, which has both positive and negative values, in contrast to the absolute return correlation matrix (not shown), which would only have large positive values. For example, the correlation between the 0–1-year T-bond active return and the 1–3-year T-bond active return in Exhibit 12 is *positive* 0.49, showing that these shorter-maturity active returns tend to move together. However, the correlation between the 0–1-year active return and the 7–10-year active return is *negative*, at −0.49, showing that these two diverse maturity active returns tend to move apart.

Exhibit 12: US Treasury Bond Estimated Active Return Risk and Correlations (return statistics from 2009 to 2018)

	Treas. 0–1	Treas. 1–3	Treas. 3–7	Treas. 7–10	Treas. 10–20
Active Vol.	3.45%	2.85%	1.05%	2.40%	4.57%

Active Corr.	Treas. 0–1	Treas. 1–3	Treas. 3–7	Treas. 7–10	Treas. 10–20
Treas. 0–1	1.000	0.49	0.21	−0.49	−0.47
Treas. 1–3	0.49	1.000	0.26	−0.49	−0.49
Treas. 3–7	0.21	0.26	1.000	−0.19	−0.33
Treas. 7–10	−0.49	−0.49	−0.19	1.000	0.46
Treas. 10–20	−0.47	−0.49	−0.33	0.46	1.000

The breadth associated with the risk estimates in Exhibit 12 is 9.4, even though there are only 5 assets. The breadth is different from the number of assets because the off-diagonal values in the correlation matrix are substantially different from zero. Exhibit 13 shows the fundamental law calculations for two sets of scores given an active portfolio risk target of 1.0% per year. The first set of scores has positive values for the shorter-maturity bonds and negative scores for the longer-maturity bonds. The associated active returns are calculated using the Grinold rule in Equation 10 and an assumed information coefficient of 0.20; for example, the active return for 10–20-year T-bonds is 0.20 × 4.57% × −1.76 = −1.61%. The active weights in Exhibit 13 are calculated by an optimizer given the constraint on active risk of 1.00%. For example, the active weight for the 10–20-year T-bonds is −19.1%, shown in the upper half of Exhibit 13. Given the benchmark weights of 20% for each asset, this results in a total weight of only 20 − 19.1 = 0.9% in the managed portfolio.

Although the information coefficient used to scale the active returns was 0.20, the first set of scores in Exhibit 13 does not represent a very ambitious forecast, so the information coefficient used in the fundamental law calculation is 0.12. The intuition for the large downward adjustment in the information coefficient is that the positive scores for the shorter-maturity bonds and the negative scores for the longer-maturity bonds are all based on essentially one active decision that interest rates will rise. Specifically, the expected active (i.e., benchmark relative) return for the managed fixed-income portfolio is $E(R_A) = (IC)\sqrt{BR}\,\sigma_A = 0.12 \times (9.4)^{1/2} \times 1.00 = 37$ bps a year.

Exhibit 13: Signals and Weights for a Fixed-Income Portfolio with Breadth of 9.4 and Active Risk of 1.00% (return statistics from 2009 to 2018)

	Treas. 0–1	Treas. 1–3	Treas. 3–7	Treas. 7–10	Treas. 10–20	IC	Active Ret.
Score	0.63	0.67	0.92	−0.46	−1.76	0.12	0.37%
Active Ret.	0.43%	0.38%	0.19%	−0.22%	−1.61%		
Active Wgt.	−1.6%	−2.1%	15.4%	7.4%	−19.1%		
Total Wgt.	18.4%	17.9%	35.4%	27.4%	0.9%		
Score	−0.22	1.20	0.23	0.57	−1.77	0.18	0.55%
Active Ret.	−0.15%	0.68%	0.05%	0.27%	−1.62%		
Active Wgt.	−11.3%	17.0%	−12.8%	24.3%	−17.2%		
Total Wgt.	8.7%	37.0%	7.2%	44.3%	2.8%		

In contrast, the second set of scores in Exhibit 13 is a more ambitious set of active forecasts that specify a modification in the shape of the yield curve. As a result, the information coefficient is 0.18, not much lower than the 0.20 value used to scale the active returns, and the expected active return for the portfolio using the fundamental law is $E(R_A) = (IC)\sqrt{BR}\,\sigma_A = 0.18 \times (9.4)^{1/2} \times 1.00 = 55$ bps a year. The fundamental law in terms of the expected information ratio for the second set of scores in Exhibit 13 is $IR = (IC)\sqrt{BR} = 0.18 \times (9.4)^{1/2} = 0.55$, alternatively calculated as the expected active return over active risk, 55/100 = 0.55.

At this relatively high information ratio, the investor may be inclined to increase the active risk to, say, 2.00% instead of 1.00%. However, given that the longest-maturity asset has a total weight that is approaching zero (i.e., 2.8%, as shown in the lower right-hand corner of Exhibit 13), such a strategy would likely require shorting; if short sells are not allowed, the transfer coefficient would likely end up being less than 1.00.

EXAMPLE 9

Breadth and Rebalancing in Active Management Strategies

Consider an active portfolio management strategy that involves decisions on overweighting or underweighting four individual assets. For example, the assets might be ETFs for four country equity markets or four different fixed-income ETFs. The active returns to Assets #1 and #2 are positively correlated, as are the active returns to Assets #3 and #4. However, the assumed risk model for active returns has no other non-zero correlations. The correlation structure in this risk model is shown in the following 4-by-4 correlation matrix, and the breadth calculation is BR = 3.2. For simplicity, we will assume that the portfolio management decisions are dichotomous; thus, each year the investor forecasts two of the assets to outperform the benchmark and the other two assets to underperform.

Correlations	#1	#2	#3	#4
#1	1.00	0.25	0.00	0.00
#2	0.25	1.00	0.00	0.00
#3	0.00	0.00	1.00	0.25
#4	0.00	0.00	0.25	1.00

1. Conceptually speaking (i.e., exact numbers are not necessary), why is the breadth less than the number of assets for this strategy?

Solution:

According to the risk model, the active returns to Assets #1 and #2 tend to move together, with a correlation coefficient of 0.25, as do the active returns for Assets #3 and #4. As a result, the 3.2 breadth of this strategy is lower than the number of assets, $N = 4$.

2. Suppose the investor predicts that Assets #1 and #2 will outperform and that Assets #3 and #4 will underperform. Conceptually speaking (i.e., exact numbers are not necessary), how will these scores affect the information coefficient in the fundamental law compared with a prediction that Assets #1 and #3 will outperform and Assets #2 and #4 will underperform?

Solution:

According to the risk model, the active returns to Assets #1 and #2 tend to move together, so a forecast that both will outperform is not as ambitious as a forecast that one will outperform while the other underperforms. As a result, the information coefficient will be adjusted downward by more under the first set of forecasts than under the second set of forecasts.

3. Suppose the active investor rebalances monthly instead of just once a year. Explain how this would affect the information ratio of this strategy, clearly stating your assumptions.

Solution:

Rebalancing monthly instead of annually could increase the breadth by a factor of 12 but only if the active management decisions for each asset are truly uncorrelated over time. For example, the breadth could increase to as much as $12 \times 3.2 = 38.4$. However, to increase the information ratio, one would have to assume that the information coefficient remains at the same level and that there are no constraints to fully implementing the active management decisions (i.e., a transfer coefficient of 1.00). For example, turnover constraints might limit the degree to which the monthly active management decisions could be fully implemented into new active positions, resulting in a lower transfer coefficient.

In summary, these examples illustrate how the information coefficient, IC, measures the strength of the return-forecasting process, or signal. The information coefficient is the correlation between the forecasted and realized security active returns and is anticipated to be positive or active management is not justified. Breadth, BR, measures the number of independent decisions made by the investor each year and is equal to the number of securities if the active returns are cross-sectionally uncorrelated. Similarly, breadth increases with the number of rebalancing periods but only if the active returns are uncorrelated over time.

Like the information coefficient, the transfer coefficient, TC, is a simple multiplicative factor in the fundamental law. It measures the extent to which constraints reduce the expected value added of the investor's forecasting ability. In the absence of constraints, the transfer coefficient is approximately 1.00, resulting in the basic form of the fundamental law. However, in practice, investors often work under constraints that result in TC values between 0.20 and 0.80. The lower transfer coefficient suggests that average performance in practice is only a fraction (20%–80%) of what would otherwise be predicted by the basic form of the fundamental law.

9 PRACTICAL LIMITATIONS

☐ describe the practical strengths and limitations of the fundamental law of active management

The limitations of the fundamental law include both practical considerations, such as ignoring transaction costs and taxes, and more conceptual issues, such as dynamic implementation over time. In this section, we focus on two limitations: the *ex ante* measurement of skill using the information coefficient and assumptions of independence in forecasts across assets and over time. The fundamental law extends the mean–variance-optimization approach to relative performance and hence has many of the same limitations of mean–variance optimization. In our discussion, we do not deal with the shortcomings of mean−variance optimization in general (e.g., assumptions of normality in return distributions or the degree of risk aversion) or the technical problems associated with the estimation and use of a risk model (e.g., the correct set of risk factors, nonlinearities, and non-stationary returns). The fundamental law takes as given that mean−variance optimization to balance risk and return against a benchmark is the correct objective function and that the investor has a way to adequately model risk.

Ex Ante Measurement of Skill

A core element of the fundamental law is the information coefficient, generally defined as the correlation between the portfolio investor's forecasts and actual outcomes. Active investors assume that the financial market they are trading in is not perfectly efficient in terms of public information and that they have some differential skill in competing with other active investors; otherwise, active management is generally not justified. Behaviorally, one might argue that investors tend to overestimate their own skills as embedded in the assumed IC, but even if that bias did not exist, questions about assessing an accurate level of skill remain. Furthermore, forecasting ability probably differs among different asset segments and varies over time.

For example, Qian and Hua (2004) expanded the basic form of the fundamental law by including the uncertainty about the level of skill, or the reality that the realized information coefficient can vary over time. Specifically, they showed that realized active portfolio risk, σ_A, is a product of both the benchmark tracking risk predicted by the risk model, denoted σ_{RM}, and the additional risk induced by the uncertainty of the information coefficient, denoted σ_{IC}:

$$\sigma_A = \sigma_{IC}\sqrt{N}\sigma_{RM}. \quad (17)$$

Their insight about "strategy risk" is derived under the simplifying assumptions that portfolio positions are unconstrained, TC = 1.00, and that breadth is the number of securities, BR = N, but can be expanded to include both refinements. In other words, they suggest that a more accurate representation of the basic fundamental law using the expression in Equation 17 is

$$\mathrm{E}\left(R_A\right) = \frac{\mathrm{IC}}{\sigma_{IC}}\sigma_A. \quad (18)$$

The key impact of accounting for the uncertainty of skill is that actual information ratios are substantially lower than predicted by an objective application of the original form of the fundamental law. Specifically, security (i.e., individual stock) selection strategies can be analytically and empirically confirmed to be 45%–91% of original estimates using the fundamental law. Like the refinement for implementation issues associated with constraints as measured by the transfer coefficient, strategy risk reduces expected and average realized information ratios. The higher the uncertainty about forecasting ability, the smaller the likely expected value added.

Independence of Investment Decisions

As we have discussed, the number of individual assets, N, is not an adequate measure of strategy breadth, BR, when the active returns between individual assets are correlated, as defined by the risk model, and forecasts are not independent from period to period. Specifically, decisions to overweight all the stocks in a given industry or all the countries in a given region because they are responding to similar influences cannot be counted as completely independent decisions, so breadth in these contexts is lower than the number of assets. Similarly, when fundamental law concepts are applied to hedging strategies using derivatives or other forms of arbitrage, breadth can increase well beyond the number of securities.

For example, arbitrage of just two securities—say, a country equity market ETF traded on two different exchanges—can have extremely high breadth (i.e., the expected active return on the strategy is large compared with the active risk). To illustrate, Clarke, de Silva, and Thorley (2006) showed that a practical measure of breadth is

$$BR = \frac{N}{1 + (N-1)\rho}, \quad (19)$$

where ρ is the same correlation coefficient in all the off-diagonal elements of the risk model. For just two securities, $N = 2$, and a correlation coefficient associated with near-arbitrage opportunities, $\rho = -0.8$, breadth could be $BR = 2/[1 - (2 - 1)0.8] = 10.0$ so that information ratios are quite high for even modest values of IC or forecasting skill.

Another example of the limitation of the fundamental law due to the lack of decision independence is the active management of fixed-income portfolios. Most descriptions of the fundamental law are based on individual stock selection strategies where the risk of equity securities is decomposed into systematic and idiosyncratic factors by a risk model. Once the systematic risk factors are removed, the active asset returns (defined as the idiosyncratic returns) are essentially independent, so breadth can be more easily determined. In contrast, almost all bonds represent some form of duration risk, as well as credit risk and optionality, so returns are highly correlated in more subtle ways. In addition, the implicit assumption of normality in the realized return distribution of bonds with default risk and embedded options is clearly unwarranted.

The limitation of independent decisions within the fundamental law also affects time-series implementation. In particular, increasing the rebalancing frequency may increase the realized information ratio but only to the extent that sequential active return forecasts are independent from period to period. Refinements on the concept of breadth—for example, Buckle (2004)—have improved the cross-sectional operationalization of the fundamental law, but more work is needed to provide conceptually useful modifications of the fundamental law in a multi-period, multi-asset setting.

In summary, the fundamental law is a useful conceptual framework in many active management applications and can even produce operational measurements of the essential elements of an active management strategy. But an understanding of the limitations of the law is warranted—particularly the issues of uncertainty in the level of assumed skill and the measurement of breadth in the face of time-dependent rebalancing policies and multi-period optimization.

EXAMPLE 10

Limitations of the Fundamental Law

1. Consider an active portfolio management strategy of selecting individual stocks in the S&P 500 on a monthly basis. The investor does a quick calculation of the fundamental law based on an information coefficient of IC = 0.05 and BR = 12 × 500 = 6,000, giving an astounding information ratio of IR = 3.87. In other words, at an active portfolio risk of 3.0%, the expected active return would be 3.87(3.0) = 11.6%.

   Provide at least two different explanations of *why* the information ratio in this example could be too high based on practical limitations of the fundamental law.

Solution:

Potential answers include the following:

1. Cross-sectional dependence: The active returns on the 500 stocks in the S&P 500 are probably correlated, so the number of independent monthly decisions is lower than 500. For example, the investor could be forecasting outperformance of all the stocks in a given industrial sector and underperformance of all the stocks in another sector.
2. Time-series dependence: The decisions on any particular stock may be correlated from month to month. For example, the forecasting process might be based on the earnings yield (reported EPS over price), which changes slowly over time. A stock that is forecasted to outperform in one month is likely to retain the outperformance forecast for several months in a row.
3. Uncertainty: Although an information coefficient of 0.05 appears to be modest, the basic form of the fundamental law does not account for uncertainty in the information coefficient or the likelihood that the information coefficient changes over time and could be different for different sets of stocks.
4. Constraints: An answer that involves accounting for such constraints as long only or turnover limits using a transfer coefficient is a weaker answer because the impact of constraints and the transfer coefficient is a well-known refinement of the fundamental law, even though it does not appear to be used in this example.

SUMMARY

We have covered a number of key concepts and principles associated with active portfolio management. Active management is based on the mathematics and principles of risk and return from basic mean–variance portfolio theory but with a focus on value added compared with a benchmark portfolio. Critical concepts include the following:

- Value added is defined as the difference between the return on the managed portfolio and the return on a passive benchmark portfolio. This difference in returns might be positive or negative after the fact but would be expected to be positive before the fact or active management would not be justified.
- Value added is related to active weights in the portfolio, defined as differences between the various asset weights in the managed portfolio and their weights in the benchmark portfolio. Individual assets can be overweighted (have positive active weights) or underweighted (have negative active weights), but the complete set of active weights sums to zero.
- Positive value added is generated when positive-active-weight assets have larger returns than negative-active-weight assets. By defining individual asset active returns as the difference between the asset total return and the benchmark return, value added is shown to be positive if and only if end-of-period realized active asset returns are positively correlated with the active asset weights established at the beginning of the period.
- Value added can come from a variety of active portfolio management decisions, including security selection, asset class allocation, and even further decompositions into economic sector weightings and geographic or country weights.
- The Sharpe ratio measures reward per unit of risk in absolute returns, whereas the information ratio measures reward per unit of risk in benchmark relative returns. Either ratio can be applied *ex ante* to expected returns or *ex post* to realized returns. The information ratio is a key criterion on which to evaluate actively managed portfolios.
- Higher information ratio portfolios can be used to create higher Sharpe ratio portfolios. The optimal amount of active management that maximizes a portfolio's Sharpe ratio is positively related to the assumed forecasting accuracy or *ex ante* information coefficient of the active strategy.
- The active risk of an actively managed strategy can be adjusted to its desired level by combining it with a position in the benchmark. Furthermore, once an investor has identified the maximum Sharpe ratio portfolio, the total volatility of a portfolio can be adjusted to its desired level by combining it with cash (two-fund separation concept).
- The fundamental law of active portfolio management began as a conceptual framework for evaluating the potential value added of various investment strategies, but it has also emerged as an operational system for measuring the essential components of those active strategies.
- Although the fundamental law provides a framework for analyzing investment strategies, the essential inputs of forecasted asset returns and risks still require judgment in formulating the expected returns.
- The fundamental law separates the expected value added, or portfolio return relative to the benchmark return, into the basic elements of the strategy:
  - *skill* as measured by the information coefficient,

- *structuring* of the portfolio as measured by the transfer coefficient,
- *breadth* of the strategy measured by the number of independent decisions per year, and
- *aggressiveness* measured by the benchmark tracking risk.

The last three of these four elements may be beyond the control of the investor if they are specified by investment policy or constrained by regulation.

- The fundamental law has been applied in settings that include the selection of country equity markets in a global equity fund and the timing of credit and duration exposures in a fixed-income fund.
- The fundamental law of active management has limitations, including uncertainty about the *ex ante* information coefficient and the conceptual definition of breadth as the number of independent decisions by the investor.

REFERENCES

Black, Fischer and Robert Litterman. 1992. "Global Portfolio Optimization." *Financial Analysts Journal* 48 (5): 28–43. 10.2469/faj.v48.n5.28

Buckle, David. 2004. "How to Calculate Breadth: An Evolution of the Fundamental Law of Active Portfolio Management." *Journal of Asset Management* 4 (6): 393–405. 10.1057/palgrave.jam.2240118

Clarke, Roger, Harindra de Silva, and Steven Thorley. 2002. "Portfolio Constraints and the Fundamental Law of Active Management." *Financial Analysts Journal* 58 (5): 48–66. 10.2469/faj.v58.n5.2468

Clarke, Roger, Harindra de Silva, and Steven Thorley. 2005. "Performance Attribution and the Fundamental Law." *Financial Analysts Journal* 61 (5): 70–83. 10.2469/faj.v61.n5.2758

Clarke, Roger, Harindra de Silva, and Steven Thorley. 2006. "The Fundamental Law of Active Portfolio Management." *Journal of Investment Management* 4 (3): 54–72.

Cremers, K.J. Martijn and Antti Petajisto. 2009. "How Active Is Your Fund Manager?" *Review of Financial Studies* 22 (9): 3329–65. 10.1093/rfs/hhp057

Fischer, Bernd and Russell Wermers. 2013. *Performance Evaluation and Attribution of Security Portfolios*. Oxford, UK: Elsevier Inc.

Grinold, Richard C. 1989. "The Fundamental Law of Active Management." *Journal of Portfolio Management* 15 (3): 30–37. 10.3905/jpm.1989.409211

Grinold, Richard C. 1994. "Alpha Is Volatility Times IC Times Score, or Real Alphas Don't Get Eaten." *Journal of Portfolio Management* 20 (4): 9–16. 10.3905/jpm.1994.409482

Grinold, Richard C. and Ronald N. Kahn. 1999. *Active Portfolio Management: A Quantitative Approach for Providing Superior Returns and Controlling Risk*. 2nd ed.New York: McGraw-Hill.

Markowitz, Harry M. 1952. "Portfolio Selection." *Journal of Finance* 7 (1): 77–91.

Qian, Edward and Ronald Hua. 2004. "Active Risk and Information Ratio." *Journal of Investment Management* 2 (3): 20–34.

Sharpe, William F. 1964. "Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk." *Journal of Finance* 19 (3): 425–42.

Treynor, J. and Fischer Black. 1973. "How to Use Security Analysis to Improve Portfolio Selection." *Journal of Business* 46:66–86. 10.1086/295508

PRACTICE PROBLEMS

1. Wei Liu makes two statements about active portfolio management:

   Statement 1 The "active return" of an actively managed portfolio is the difference between the portfolio's return and the return on the benchmark portfolio, and it is equal to the managed portfolio's alpha.

   Statement 2 The active weights are the differences in the managed portfolio's weights and the benchmark's weights.

   Are Liu's statements correct?

   A. Only Statement 1 is correct.

   B. Only Statement 2 is correct.

   C. Both statements are correct.

2. The benchmark weights and returns for each of the five stocks in the Capitol Index are given in the following table. The Tukol Fund uses the Capitol Index as its benchmark, and the fund's portfolio weights are also shown in the table.

Stock	Portfolio Weight (%)	Benchmark Weight (%)	20X2 Return (%)
1	30	24	14
2	30	20	15
3	20	20	12
4	10	18	8
5	10	18	10

   What is the value added (active return) for the Tukol Fund?

   A. 0.00%

   B. 0.90%

   C. 1.92%

3. Consider the following asset class returns for calendar year 20X2:

Asset Class	Portfolio Weight (%)	Benchmark Weight (%)	Portfolio Return (%)	Benchmark Return (%)
Domestic equities	55	40	10	8
International equities	20	30	10	9
Bonds	25	30	5	6

   What is the value added (or active return) for the managed portfolio?

   A. 0.25%

   B. 0.35%

   C. 1.05%

The following information relates to questions 4-10

James Frazee is chief investment officer at H&F Capital Investors. Frazee hires a third-party adviser to develop a custom benchmark for three actively managed balanced funds he oversees: Fund X, Fund Y, and Fund Z. (Balanced funds are funds invested in equities and bonds.) The benchmark needs to be composed of 60% global equities and 40% global bonds. The third-party adviser submits the proposed benchmark to Frazee, who rejects the benchmark based on the following concerns:

Concern 1: Many securities he wants to purchase are not included in the benchmark portfolio.

Concern 2: One position in the benchmark portfolio will be somewhat costly to replicate.

Concern 3: The benchmark portfolio is a float-adjusted, capitalization-weighted portfolio.

After the third-party adviser makes adjustments to the benchmark to alleviate Frazee's concerns, Frazee accepts the benchmark portfolio. He then asks his research staff to develop risk and expected return forecasts for Funds X, Y, and Z as well as for the benchmark. The forecasts are presented in Exhibit 1.

Exhibit 1: Forecasted Portfolio Statistics for Funds X, Y, and Z and the Benchmark

	Fund X	Fund Y	Fund Z	Benchmark
Portfolio weights:				
Global equities (%)	60.0	65.0	68.0	60.0
Global bonds (%)	40.0	35.0	32.0	40.0
Expected return (%)	10.0	11.6	13.2	9.4
Expected volatility (%)	17.1	18.7	22.2	16.3
Active risk (%)	5.2	9.2	15.1	N/A
Sharpe ratio (SR)	0.45	0.50	0.49	0.44

Note: Data are based on a risk-free rate of 2.3%.

Frazee decides to add a fourth offering to his group of funds, Fund W, which will use the same benchmark as in Exhibit 1. Frazee estimates Fund W's information ratio to be 0.35. He is considering adding the following constraint to his portfolio construction model: Fund W would now have maximum over- and underweight constraints of 7% on single-country positions.

Frazee conducts a search to hire a manager for the global equity portion of Fund W and identifies three candidates. He asks the candidates to prepare risk and return forecasts relative to Fund W's benchmark based on their investment strategy, with the only constraint being no short selling. Each candidate develops independent annual forecasts with active return projections that are uncorrelated and constructs a portfolio made up of stocks that are diverse both geographically and across economic sectors. Selected data for the three candidates' portfolios are presented in Exhibit 2.

Exhibit 2: Forecasted Portfolio Data for Equity Portion of Fund W

	Candidate A	Candidate B	Candidate C
Rebalancing	Annually	Annually	Annually
Number of securities	100	64	36
Information ratio (IR)	0.582	0.746	0.723
Transfer coefficient (TC)	0.832	0.777	0.548
Information coefficient*	0.07	0.12	0.22

** Information coefficient based on previously managed funds.*

Frazee asks Candidate C to re-evaluate portfolio data given the following changes:

Change 1: Fix the number of securities to 50.

Change 2: Rebalance on a semi-annual basis.

Change 3: Add maximum over- or underweight constraints on sector weightings.

4. Which of Frazee's concerns *best* justifies his decision to reject the proposed benchmark?

   A. Concern 1

   B. Concern 2

   C. Concern 3

5. Based on Exhibit 1, the expected active return from asset allocation for Fund X is:

   A. negative.

   B. zero.

   C. positive.

6. Based on Exhibit 1, which fund is expected to produce the greatest consistency of active return?

   A. Fund X

   B. Fund Y

   C. Fund Z

7. Based on Exhibit 1, combining Fund W with a fund that replicates the benchmark would produce a Sharpe ratio *closest* to:

   A. 0.44.

   B. 0.56.

   C. 0.89.

8. If Frazee added the assumption he is considering in Fund W's portfolio construc-

tion, it would *most likely* result in:

A. a decrease in the optimal aggressiveness of the active strategy.

B. the information ratio becoming invariant to the level of active risk.

C. an increase in the transfer of active return forecasts into active weights.

9. Based on the data presented in Exhibit 2, the candidate with the greatest skill at achieving active returns appears to be:

A. Candidate A.

B. Candidate B.

C. Candidate C.

10. Which proposed change to Fund W would *most likely* decrease Candidate C's information ratio?

A. Change 1

B. Change 2

C. Change 3

The following information relates to questions 11-14

John Martinez is assessing the performance of the actively managed diversified asset portfolio. The diversified asset portfolio is invested in equities, bonds, and real estate, and allocations to these asset classes and to the holdings within them are unconstrained.

Selected return and financial data for the portfolio for 2019 are presented in Exhibit 1.

Exhibit 1: Diversified Asset Portfolio 2019 Portfolio Performance

	Sub-Portfolio Return (%)	Benchmark Return (%)	Portfolio Allocation (%)	Strategic Asset Allocation (%)
Equities sub-portfolio	36.9	31.6	63	60
Bond sub-portfolio	−2.4	−2.6	28	35
Real estate sub-portfolio	33.4	28.3	9	5

Martinez uses several risk-adjusted return metrics to assess the performance of the diversified asset portfolio, including the information ratio and the Sharpe ratio. Selected risk, return, and statistical data for the portfolio are presented in Exhibit 2.

Exhibit 2: Diversified Asset Portfolio Data, 2000-2019

	Transfer Coefficient (TC)	Information Coefficient (IC)	Breadth (BR)
Equities sub-portfolio	0.90	0.091	21
Bond sub-portfolio	0.79	0.087	23
Real estate sub-portfolio	0.86	0.093	19

Martinez has recently hired Kenneth Singh to help him evaluate portfolios. Martinez asks Singh about the possible effects on the portfolio's information ratio if cash were added to the diversified asset portfolio or if the aggressiveness of the portfolio's active weights were increased. Singh responds with two statements:

Statement 1 Adding cash to the portfolio would change the portfolio's information ratio.

Statement 2 Increasing the aggressiveness of active weights would not change the portfolio's information ratio.

11. Based on Exhibit 1, the value added to the diversified asset portfolio attributable to the security selection decision in 2019 was *closest* to:

A. 2.3%.

B. 3.9%.

C. 6.1%.

12. Based on Exhibit 1, the value added of the diversified asset portfolio attributable to the asset allocation decision in 2019 was *closest* to:

A. 2.3%.

B. 3.9%.

C. 6.1%.

13. Based on data in Exhibit 2 and using the information ratio as the criterion for evaluating performance, which sub-portfolio had the best performance in the period 2000–2019?

A. The bond sub-portfolio.

B. The equities sub-portfolio.

C. The real estate sub-portfolio.

14. Which of Singh's statements regarding the information ratio is correct?

A. Only Statement 1

B. Only Statement 2

C. Both Statement 1 and Statement 2

15. Gertrude Fischer mentions two properties of the Sharpe ratio and the information ratio that she says are very useful.

Property 1	The Sharpe ratio is unaffected by the addition of cash or leverage in a portfolio.
Property 2	The information ratio for an unconstrained portfolio is unaffected by the aggressiveness of the active weights.

Are Fischer's two properties correct?

A. Yes.

B. No. Only Property 1 is correct.

C. No. Only Property 2 is correct.

16. An analyst is given the following information about a portfolio and its benchmark. In particular, the analyst is concerned that the portfolio is a closet index fund. The T-bill return chosen to represent the risk-free rate is 0.50%.

	Benchmark	Portfolio
Return	8.75%	8.90%
Risk	17.50%	17.60%
Active return	0.00%	0.15%
Active risk	0.00%	0.79%
Sharpe ratio	0.4714	0.4773
Information ratio	N/A	0.1896

Which of the following three statements *does not* justify your belief that the portfolio is a closet index?

i. The Sharpe ratio of the portfolio is close to the Sharpe ratio of the benchmark.
ii. The information ratio of the portfolio is relatively small.
iii. The active risk of the portfolio is very low.

A. Statement I

B. Statement II

C. Statement III

17. You have a portfolio 100% allocated to a manager with an *ex post*, active risk at 8.0%. You choose to allocate a 75% position to the active manager and 25% to the benchmark to bring the portfolio back to your target active risk of 6.0%. If the manager's information ratio is 0.50, what happens to the information ratio of the portfolio after the reallocation?

A. The information ratio increases because the lower active risk reduces the denominator of the ratio.

B. The information ratio remains unchanged because allocations between the active portfolio and the benchmark don't affect the information ratio.

C. The information ratio decreases because allocating some of the portfolio to the benchmark means that the external manager generates less active return.

The following information relates to questions 18-19

	S&P 500	Indigo Fund
Expected annual return	9.0%	10.5%
Return standard deviation	18.0%	25.0%
Sharpe ratio	0.333	0.30
Active return		1.2%
Active risk		8.0%
Information ratio		0.15

Note: Data are based on a risk-free rate of 2.3%.

18. What is the maximum Sharpe ratio that a manager can achieve by combining the S&P 500 benchmark portfolio and the Indigo Fund?

A. 0.333

B. 0.365

C. 0.448

19. Which of the following pairs of weights would be used to achieve the highest Sharpe ratio and optimal amount of active risk through combining the Indigo Fund and benchmark portfolio, respectively?

A. 1.014 on Indigo and −0.014 on the benchmark

B. 1.450 on Indigo and −0.450 on the benchmark

C. 1.500 on Indigo and −0.500 on the benchmark

20. The benchmark portfolio is the S&P 500. Which of the following three portfolios can be combined with the benchmark portfolio to produce the highest combined Sharpe ratio?

	S&P 500	Portfolio A	Portfolio B	Portfolio C
Expected annual return	9.0%	10.0%	9.5%	9.0%
Return standard deviation	18.0%	20.0%	20.0%	18.0%
Sharpe ratio	0.333	0.350	0.325	0.333
Active return	0	1.0%	0.5%	0
Active risk	0	10.0%	3.0%	2.0%

Note: Data are based on a risk-free rate of 2.3%.

A. Portfolio A

B. Portfolio B

C. Portfolio C

21. You are considering three managers for a small-cap growth mandate. After careful analysis, you produce the following forward-looking expectations about the managers' active risk and active return:

	Manager A	Manager B	Manager C
Active return	0.7%	0.6%	1.2%
Active risk	3.2%	3.1%	6.3%

If you intend to rely on the information ratio to make your decision, which manager should you choose?

A. Manager A

B. Manager B

C. Manager C

22. Based on the fundamental law of active management, if a portfolio manager has an information ratio of 0.75, an information coefficient of 0.1819, and a transfer coefficient of 1.0, how many securities are in the portfolio manager's fund, making the assumption that the active returns are uncorrelated.

A. About 2

B. About 4

C. About 17

23. Two analysts make the following statements about the transfer coefficient in the expanded fundamental law of active management:

Analyst One says, "The transfer coefficient measures how well the realized returns correlate with the anticipated returns, adjusted for risk."

Analyst Two says, "The transfer coefficient measures how well the realized returns correlate with the active weights, adjusted for risk."

Which, if either, analyst is correct?

A. Only Analyst One is correct.

B. Only Analyst Two is correct.

C. Neither analyst is correct.

24. The expanded fundamental law of active management is stated as follows:

$$E\left(R_A\right) = (TC)\,(IC)\,\sqrt{BR}\,\sigma_A.$$

Which component on the righthand side represents the extent to which the portfolio manager's expectations are realized? The

A. transfer coefficient, TC.

B. information coefficient, IC.

C. breadth, BR.

The following information relates to questions 25-26

You are analyzing three investment managers for a new mandate. The following table provides the managers' ex-ante active return expectations and portfolio weights. The last two columns include the risk and the *ex post*, realized active returns for the four stocks. Use the following data for the following two questions:

	Manager 1		Manager 2		Manager 3			
	Δw	$E(R_A)$	Δw	$E(R_A)$	Δw	$E(R_A)$	Risk	Realized R_A
Security 1	−0.125	0.03	0.2	0.04	−0.05	0.025	0.17	0.06
Security 2	0.025	0.04	0	0.01	0.05	0.015	0.10	0.07
Security 3	0.075	0.05	−0.1	0	0.05	0.005	0.12	0.04
Security 4	0.025	0.06	−0.1	0.02	−0.05	0.015	0.25	0.02

25. Suppose all three managers claim to be good at forecasting returns. According to the expanded fundamental law of active management, which manager is the best at efficiently building portfolios by anticipating future returns?

A. Manager 1

B. Manager 2

C. Manager 3

26. Suppose all three managers claim to be efficient in portfolio construction. According to the expanded fundamental law of active management, which manager is the best at building portfolios to make full use of their ability to correctly anticipate returns?

A. Manager 1

B. Manager 2

C. Manager 3

27. Manager 1 has an information coefficient of 0.15, a transfer coefficient of 1.0, and invests in 50 securities. Manager 2 has a different strategy, investing in more securities; however, he is subject to investment constraints that reduce his transfer coefficient. Manager 2 has an information coefficient of 0.10, a transfer coefficient of 0.8, and invests in 100 securities. The investment selections of each manager are independent decisions. If both managers target an active risk of 5.0%, which manager will have the greater expected active return?

A. Manager 1

B. Manager 2

C. Both managers will have the same active return.

28. Nick Young is concerned that Goudon Partners, one of his money managers, overestimates its expected active return because Goudon overstates its strategy breadth. Young makes two notes about his concern:

    Note 1 Although Goudon claims that the number of independent asset decisions is high because it uses 200 stocks, many of these stocks cluster in industries where the same general analysis applies to several stocks.

    Note 2 Goudon claims that each stock is independent and evaluated each month, or 12 times per year. These analyses are not independent because some of their strategies, such as favoring a particular industry or favoring value stocks, persist beyond one month. For example, a strategy of favoring low-P/E stocks will persist for several months and the investment decisions are not independent.

    If his judgments are correct, are Young's notes about the overstatement of breadth correct?

    A. Only Note 1 is correct.

    B. Only Note 2 is correct.

    C. Both Notes 1 and 2 are correct.

29. Caramel Associates uses the fundamental law to estimate its expected active returns. Two things have changed. First, Caramel will lower its estimate of the information coefficient because they felt their prior estimates reflected overconfidence. Second, their major clients have relaxed several constraints on their portfolios—including social screens, prohibitions on short selling, and constraints on turnover. Which of these changes will increase the expected active return?

    A. Only the lower information coefficient.

    B. Only the relaxation of several portfolio constraints.

    C. Both the lower information coefficient and the relaxation of portfolio constraints.

SOLUTIONS

1. B is correct. Although the first part of Statement 1 is correct (active return, or value added, equals the difference between the managed portfolio return and the benchmark return), active return is not the same as alpha. In other words, $R_A = R_P - R_B$, while $\alpha_P = R_P - \beta_P \times R_B$. Statement 2 correctly defines active weights.

2. B is correct. The portfolio active return is equal to the portfolio return minus the benchmark return:

   $R_A = R_P - R_B$.

   The portfolio return is $R_P = \sum_{i=1}^{n} w_{P,i} R_i$;

   $R_P = 0.30(14\%) + 0.30(15\%) + 0.20(12\%) + 0.10(8\%) + 0.10(10\%) = 12.9\%$.

   The benchmark return is $R_B = \sum_{i=1}^{n} w_{B,i} R_i$;

   $R_B = 0.24(14\%) + 0.20(15\%) + 0.20(12\%) + 0.18(8\%) + 0.18(10\%) = 12.0\%$.

   Thus, the active return is

   $R_A = R_P - R_B = 12.9\% - 12.0\% = 0.9\%$.

   Note that this same correct answer can be obtained in two other equivalent ways. The active weights are the differences between the portfolio and benchmark weights, or $\Delta w_i = w_{P,i} - w_{B,i}$. Computing the active weights from the table provided, the active return is

   $$R_A = \sum_{i=1}^{N} \Delta w_i R_i$$
   $$= 0.06(14\%) + 0.10(15\%) + 0(12\%) - 0.08(8\%) - 0.08(10\%)$$
   $$= 0.9\%.$$

   Finally, we could express the active security returns as their differences from the benchmark return, or $R_{Ai} = R_i - R_B$. Computing the active security returns from the table provided, the portfolio active return is the sum product of the active weights and the active security returns:

   $$R_A = \sum_{i=1}^{N} \Delta w_i R_{Ai}$$
   $$= 0.06(2\%) + 0.10(3\%) + 0(0\%) - 0.08(-4\%) - 0.08(-2\%)$$
   $$= 0.9\%.$$

3. C is correct. The active return is equal to the portfolio return minus the benchmark return:

   $$R_A = R_P - R_B = \sum_{j=1}^{M} w_{P,j} R_{P,j} - \sum_{j=1}^{M} w_{B,j} R_{B,j}.$$

   The portfolio return is $R_P = \sum_{i=1}^{n} w_{P,i} R_i = 0.55(10\%) + 0.20(10\%) + 0.25(5\%) =$ 8.75%.

   The benchmark return is $R_B = \sum_{i=1}^{n} w_{B,i} R_i = 0.40(8\%) + 0.30(9\%) + 0.30(6\%) =$ 7.70%.

Thus, $R_A = R_P - R_B = 8.75\% - 7.70\% = 1.05\%$.

4. A is correct. Because the benchmark does not contain many assets that Frazee wants to invest in, the benchmark may not be representative of his investment approach. Concern 2, as stated, is less important because it does not imply that the cost of replicating the benchmark is a serious concern. Finally, Concern 3 actually states a generally positive feature of the benchmark.

5. B is correct. Active return from asset allocation is derived from differences between the benchmark weight and the portfolio weight across asset classes. For Fund X, the expected active return from asset allocation is calculated as

$$\sum_{j=1}^{M} \Delta w_j R_{B,j} = (60 - 60) R_{B,e} + (40 - 40) R_{B,b} = 0,$$

where Δw_j is the difference in the active portfolio and the benchmark asset weights, $R_{B,e}$ is the benchmark's return from global equities, and $R_{B,b}$ is the benchmark's return from global bonds.

Because Fund X has the same asset weights as the benchmark across the two asset classes (60% global equities, 40% global bonds), the expected active return from asset allocation is zero.

6. C is correct. The IR measures the consistency of active return. The IR is calculated for the three funds as follows:

$$\text{IR} = \frac{R_P - R_B}{\sigma(R_P - R_B)} = \frac{R_A}{\sigma_A}.$$

IR for Fund X = (10.0 − 9.4)/5.2 = 0.6/5.2 = 0.12.

IR for Fund Y = (11.6 − 9.4)/9.2 = 2.2/9.2 = 0.24.

IR for Fund Z = (13.2 − 9.4)/15.1 = 3.8/15.1 = 0.25.

Fund Z has the largest IR and thus is expected to produce the greatest consistency of active return.

7. B is correct. Given the IR for Fund W of 0.35 and the benchmark's SR of 0.44, the combination of the benchmark portfolio and Fund W would produce an SR of 0.56, calculated as follows:

$$\text{SR}_P^2 = \text{SR}_B^2 + \text{IR}^2;$$

$$\text{SR}_P = (0.44^2 + 0.35^2)^{0.5} = 0.56.$$

8. A is correct. The new assumption adds constraints to Fund W. The IR for a constrained portfolio generally decreases with the aggressiveness of the strategy because portfolio constraints reduce the transfer of active return forecasts into active weights. Furthermore, the optimal active risk is given by the following formula:

$$\sigma_A = \text{TC}\frac{\text{IR}}{\text{SR}_B}\sigma_B.$$

The addition of portfolio constraints reduces the TC, thus also reducing the optimal active risk.

So, having maximum over- and underweight constraints on single-country positions decreases the optimal aggressiveness of the active management strategy.

9. B is correct. The IR measures the consistency of active return generation. A higher ratio generally indicates better managerial skill at achieving active returns on a risk-adjusted basis. The IR for Candidate B (0.746) is higher than the IR for Candidate A (0.582) and Candidate C (0.723).

   Thus, Candidate B appears to have the greatest skill, as indicated by the highest IR of 0.746.

10. C is correct. The IR is calculated as $IR = (TC)(IC)\sqrt{BR}$, where BR is breadth. Change 3, establishing new constraints of caps on the over- and underweight of sectors, reduces the correlation of optimal active weights with the actual active weights, which results in a decreased TC and thus a decrease in the IR. Change 1 (increasing portfolio size from 36 to 50) and Change 2 (increasing the frequency of rebalancing from annually to semi-annually) would both likely have the effect of increasing the BR of the portfolio, which would increase the IR.

11. B is correct. Based on the differences in returns for the portfolio and benchmark in Exhibit 1, the value added by each asset class within the portfolio is shown in the following table:

	Sub-Portfolio Return (%)	Benchmark Return (%)	Value Added (%)	Portfolio Allocation (%)
Equities sub-portfolio	36.9	31.6	5.3	63
Bond sub-portfolio	−2.4	−2.6	0.2	28
Real estate sub-portfolio	33.4	28.3	5.1	9

   The value added from security selection is calculated as the sum of the actual portfolio weights multiplied by each sub-portfolio's value added measure. Thus, the value added from security selection is calculated as: Value added from security selection = 0.63(5.3%) + 0.28(0.2%) + 0.09(5.1%) = 3.9%.

   A is incorrect. It represents the value added from asset allocation (2.3%).

   C is incorrect. It represents the total value added (2.3% + 3.9% = 6.1%, with rounding).

12. A is correct. The value added from asset allocation is calculated as the sum of the differences in the weights between the strategic (benchmark) allocation and the actual sub-portfolio allocation multiplied by each sub-portfolio's benchmark return.

	Benchmark Return (%)	Actual Asset Allocation (%)	Strategic Asset Allocation (%)	Actual – Strategic Asset Allocation (%)
Equities sub-portfolio	31.6	63	60	+3
Bond sub-portfolio	−2.6	28	35	−7
Real estate sub-portfolio	28.3	9	5	+4

   Thus, the value added by the active asset allocation decision is calculated as

Value added from asset allocation decision = 0.03(31.6%) − 0.07(−2.6%) + 0.04(28.3%)

= 2.3%.

B is incorrect. It is the value added from security selection.

C is incorrect. It is the total value added.

13. B is correct. The information ratio for a portfolio can be expressed as follows:

$$\text{IR} = (\text{TC})(\text{IC})\sqrt{\text{BR}}.$$

The information ratios for the three sub-portfolios are calculated as follows:

	Information Ratio
Equities sub-portfolio	$0.90 \times 0.091 \times (21)^{0.5} = 0.38$
Bond sub-portfolio	$0.79 \times 0.087 \times (23)^{0.5} = 0.33$
Real estate sub-portfolio	$0.86 \times 0.093 \times (19)^{0.5} = 0.35$

Based on the information ratio, the equities sub-portfolio outperformed the real estate sub-portfolio. The information ratio for the equities sub-portfolio of 0.38 was higher than the information ratio for the real estate sub-portfolio of 0.35 and the bond sub-portfolio of 0.33.

14. C is correct. The information ratio for a portfolio of risky assets will generally shrink if cash is added to the portfolio. Because the diversified asset portfolio is an unconstrained portfolio, its information ratio would be unaffected by an increase in the aggressiveness of active weights.

15. A is correct. Both properties are correct. For Property 1, if w_P is the weight of an actively managed portfolio and $(1 - w_P)$ is the weight on risk-free cash, changing w_P does not change the Sharpe ratio, as can be seen in this equation:

$$\text{SR}_C = \frac{R_C - R_F}{\sigma_C} = \frac{w_P\left(R_P - R_F\right)}{w_P \sigma_P} = \text{SR}_P.$$

For Property 2, the information ratio of an unconstrained portfolio is unaffected by multiplying the active security weights, Δw_i, by a constant.

16. B is correct. A closet index will have a very low active risk and will also have a Sharpe ratio very close to the benchmark. Therefore, Statements I and III are consistent with a closet index portfolio. A closet index's information ratio can be indeterminate (because the active risk is so low) and is often negative due to management fees.

17. B is correct. The information ratio is unaffected by rebalancing the active portfolio and the benchmark portfolio. In this case, the active return and active risk are both reduced by 25% and the information ratio will be unchanged.

18. B is correct. The highest squared Sharpe ratio of an actively managed portfolio is

$$\text{SR}_P^2 = \text{SR}_B^2 + \text{IR}^2 = 0.333^2 + 0.15^2 = 0.1334.$$

The highest Sharpe ratio is $\text{SR}_P = \sqrt{0.1334} = 0.365$.

19. A is correct. The optimal amount of active risk is

$$\sigma_A = \frac{\text{IR}}{\text{SR}_B}\sigma_B = \frac{0.15}{0.333}18.0\% = 8.11\%.$$

The weight on the active portfolio (Indigo) would be 8.11%/8.0% = 1.014, and the weight on the benchmark portfolio would be 1 − 1.014 = −0.014.

We can demonstrate that these weights achieve the maximum Sharpe ratio (of 0.365). Note that 8.11% is the optimal level of active risk and that Indigo has an expected active return of 1.014(1.2%) = 1.217% over the benchmark and a total excess return of 6.0% + 1.217% = 7.217%. The portfolio total risk is

$\sigma_P^2 = \sigma_B^2 + \sigma_A^2 = 18.0^2 + 8.111^2 = 389.788.$

Taking the square root, $\sigma_P = 19.743$, and the optimal Sharpe ratio is indeed 7.217/19.743 = 0.365.

20. B is correct. The optimal active portfolio is the portfolio with the highest information ratio, the ratio of active return to active risk. The IRs for the three active portfolios are as follows:

    $IR_A = 1.0/10.0 = 0.10$

    $IR_B = 0.\ 5/3.0 = 0.167$

    $IR_C = 0/2.0 = 0.00$

    Portfolio B has the highest IR and is the best active portfolio; it is therefore the best portfolio to combine with the benchmark.

21. A is correct. Manager A has the highest information ratio. The information ratio is defined as $IR = \frac{\text{active return}}{\text{active risk}}$. The managers in this example have the following information ratios:

	Manager A	Manager B	Manager C
Information ratio	0.7/3.2 = 0.219	0.6/3.1 = 0.194	1.2/6.3 = 0.190

22. C is correct. Using the equation $IR^* = IC \times \sqrt{BR}$ and assuming that breadth can be interpreted as number of securities in the portfolio, solving for breadth in the equation yields $\left(\frac{0.75}{0.1819}\right)^2 = 17.000$.

23. C is correct. The transfer coefficient measures how well the anticipated (*ex ante*), risk-adjusted returns correlate with the risk-adjusted active weights. This is also expressed in the equation for the transfer coefficient: $TC = \rho(\mu_i/\sigma_i, \Delta w_i\sigma_i)$.

24. B is correct. The IC measures an investment manager's ability to forecast returns.

25. C is correct. The proper statistic to calculate is the information coefficient, and it is defined as follows:

    $$IC = \rho\left(\frac{R_{Ai}}{\sigma_i}, \frac{\mu_i}{\sigma_i}\right).$$

    A manager is a good forecaster if his or her *ex ante*, active return expectations (forecasts) are highly correlated with the realized active returns. The information coefficient requires that these forecasts and realized returns be risk-weighted. When this is done for the three managers, the risk-weighted forecasts and realized returns are:

	Risk-Weighted Forecasts, μ_i/σ_i			R_{Ai}/σ_i
	Manager 1	Manager 2	Manager 3	Realized
Security 1	0.176	0.235	0.147	0.353
Security 2	0.400	0.100	0.150	0.700

	Risk-Weighted Forecasts, μ_i/σ_i			R_{Ai}/σ_i
	Manager 1	**Manager 2**	**Manager 3**	**Realized**
Security 3	0.417	0.000	0.042	0.333
Security 4	0.240	0.080	0.060	0.080

The ICs are found by calculating the correlations between each manager's forecasts and the realized risk-weighted returns. The three managers have the following ICs:

	Manager 1	Manager 2	Manager 3
Information coefficient	0.5335	0.0966	0.6769

Manager 3 has the highest IC.

26. B is correct. The proper statistic to calculate is the transfer coefficient, and it is defined as follows:

    $TC = \rho(\mu_i/\sigma_i, \Delta w_i \sigma_i)$

    The TC is the cross-sectional correlation between the forecasted active security returns and the actual active weights, adjusted for risk.

	Risk-Weighted Forecasts, μ_i/σ_i			Risk-Adjusted Weights, $\Delta w_i \sigma_i$		
	Manager 1	**Manager 2**	**Manager 3**	**Manager 1**	**Manager 2**	**Manager 3**
Security 1	0.1765	0.2353	0.1471	−0.0213	0.0340	−0.0085
Security 2	0.4000	0.1000	0.1500	0.0025	0.0000	0.0050
Security 3	0.4167	0.0000	0.0417	0.0090	−0.0120	0.0060
Security 4	0.2400	0.0800	0.0600	0.0063	−0.0250	−0.0125

The three managers have the following TCs:

	Manager 1	Manager 2	Manager 3
Transfer coefficient	0.7267	0.8504	−0.0020

Manager 2 has the highest TC.

27. A is correct. Manager 1's IR = TC × IC × $\sqrt{BR}$ = 1.0 × 0.15 × $\sqrt{50}$ = 1.06. Manager 2's IR = 0.8 × 0.10 × $\sqrt{100}$ = 0.80. Manager 1's active return is 1.06(5.0) = 5.3%, and Manager 2's expected active return is 0.80(5.0) = 4.0%. Manager 1 has the greater expected active return.

28. C is correct. If the decisions about each of the 200 stocks are not independent, and if the decisions about a stock from one month to the next are not independent, then Goudon Partners is overstating its estimates of its breadth and its expected active returns.

29. B is correct. Although the relaxation of portfolio constraints will increase the transfer coefficient (and expected active returns), the lower information coefficient reduces the information ratio and the expected active return.

LEARNING MODULE

3

Exchange-Traded Funds: Mechanics and Applications

by Joanne M. Hill, PhD, and Dave Nadig.

Joanne M. Hill, PhD (USA). Dave Nadig is CIO and Director of Research at ETFTrends.com and etfdb.com (USA).

LEARNING OUTCOMES

Mastery	*The candidate should be able to:*
☐	explain the creation/redemption process of ETFs and the function of authorized participants
☐	describe how ETFs are traded in secondary markets
☐	describe sources of tracking error for ETFs
☐	describe factors affecting ETF bid–ask spreads
☐	describe sources of ETF premiums and discounts to NAV
☐	describe costs of owning an ETF
☐	describe types of ETF risk
☐	identify and describe portfolio uses of ETFs

INTRODUCTION

1

Exchange-traded funds (ETFs) have grown rapidly since their invention in the early 1990s, in large part because of their low associated cost, exchange access, holdings transparency, and range of asset classes available. Growth in ETFs has also been driven by the increased use of index-based investing. ETF investors need to understand how these products work and trade and how to choose from the numerous options available. Although many ETFs are organized under the same regulation as mutual fund products, there are important differences related to trading and tax efficiency. ETFs have features that can make them more tax efficient than traditional mutual funds, and not all ETFs are organized like mutual funds. ETFs can be based on derivative strategies, use leverage and shorting, and be offered in alternate structures, such as exchange-traded notes (ETNs), which have their own unique risks.

Understanding how ETF shares are created and redeemed is key to understanding how these products can add value in a portfolio. Because so many ETFs track indexes, understanding their index tracking or tracking error is also critical. Investors should also understand how to assess an ETF's trading costs, including differences between the ETF's market price and the fair value of its portfolio holdings.

We start with a discussion of the primary and secondary markets for ETFs, including the creation/redemption process, before moving on to important investor considerations, such as costs and risks. We then explain how ETFs are use in strategic, tactical, and portfolio efficiency applications.

2 ETF MECHANICS

- ☐ explain the creation/redemption process of ETFs and the function of authorized participants
- ☐ describe how ETFs are traded in secondary markets

Exchange-traded funds function differently from mutual funds because of their structure, with the key difference in an ETF's method of share creation and redemption. Mutual fund shares must be purchased or sold at the end of the day from the fund manager (or via a broker) at the closing net asset value (NAV) of the fund's holdings, in a cash-for-shares or shares-for-cash swap. In contrast, an ETF trades intraday, or during the trading day, just like a stock. ETF shares are created or redeemed in kind, in a shares-for-shares swap.

ETFs are intrinsically linked to the creation/redemption process. Creation/redemption enables ETFs to operate at lower cost and with greater tax efficiency than mutual funds and generally keeps ETF prices in line with their NAVs. Unlike stocks, which come to market via an initial public offering of fixed size, ETFs can be created or redeemed continuously. ETF transactions take place in two interrelated markets. Understanding how this mechanism works is key to understanding both the benefits and potential risks of ETFs.

The primary market for ETF trading is that which exists on an over-the-counter (OTC) basis between **authorized participants** (APs), a special group of institutional investors, and the ETF issuer, or sponsor. This process is referred to as **creation/redemption**. These primary market transactions are the only way that shares of the ETF can be created or redeemed. The "trade" in this market is in kind: A pre-specified basket of securities (which can include cash) is exchanged for a certain number of shares in the ETF.

ETF shares trade in the secondary market on exchanges. For investors, exchange trading is the only way to buy or sell ETFs. Like stocks, ETFs are bought and sold on exchanges through a brokerage account. This secondary market trading is perhaps the most novel feature of ETFs.

In-kind creation/redemption creates the unique benefits ETFs offer—as well as some of their risks. Here we explain ETFs' unique creation/redemption mechanism, the role of APs, and how the creation/redemption mechanism affects ETF design. ETF trading and settlement on primary and secondary markets is also covered.

The Creation/Redemption Process

The best way to understand the creation/redemption process is to step through the process from an investor's perspective.

Imagine you're an investor and you want to invest in an ETF. The process is simple: You place a buy order in your brokerage account the same way you would place an order to buy any publicly listed equity security, and your broker submits that order to the public market to find a willing seller: another investor or a market maker (i.e., a broker/dealer who stands ready to take the opposite side of the transaction). The order is executed, and you receive shares of the ETF in your brokerage account just as if you transacted in a stock.

At this point, the ETF manager (also referred to as the ETF issuer or sponsor) is not involved in the transaction. The ETF issuer does not know that you have bought these shares, nor does it receive an inflow of money to invest. Shares simply transfer in the open market, the secondary market for ETF shares, from one investor (the seller) to another (the buyer) and go through a settlement process based on the local exchange where the transaction took place. The process sounds simple, but if you can only buy ETF shares from another investor, where do the shares come from initially? How does money get invested into the fund?

The only investors who can create or redeem new shares of an ETF are a special group of institutional investors called *authorized participants*. APs are large broker/dealers, often market makers, who are authorized by the ETF issuer to participate in the creation/redemption process. The AP creates new ETF shares by transacting in kind with the ETF issuer. This in-kind swap happens off the exchange, in the primary market for the ETF, where APs transfer securities to (for creations) or receive securities from (for redemptions) the ETF issuer, in exchange for ETF shares. This is a prescribed, structured transaction with its own set of rules.

Each business day, the ETF manager publishes a list of required in-kind securities for each ETF. For instance, an S&P 500 Index ETF will typically list the index securities in quantities that reflect the index weighting. The list of securities specific to each ETF and disclosed publicly each day is called the **creation basket**. This basket also serves as the portfolio for determining the intrinsic net asset value of the ETF based on prices during the trading day.

To create new shares, an AP acquires the securities in the creation basket in the specified share amounts (generally by transacting in the public markets or using securities the AP happens to have in inventory). The AP then delivers this basket of securities to the ETF manager in exchange for an equal value in ETF shares. This exchange of shares happens after markets are closed through the settlement process. Importantly, the pricing of both the ETF and the basket is of minimal concern in this exchange: If the issuer receives 100 shares of a certain stock as part of the creation basket, the price the AP might have paid to acquire that stock or what its price happens to be at the end of the day is not relevant to the exchange taking place. Because it is an in-kind transaction, all that matters is that 100 shares of the required stock move from the AP's account to the ETF's account. Similarly, when the issuer delivers ETF shares to an AP, the ETF's closing NAV is not relevant.

These transactions between the AP and the ETF manager are done in large blocks called **creation units**, usually but not always equal to 50,000 shares of the ETF. This in-kind exchange involves the basket of underlying securities in exchange for a number of ETF shares of equal value.

The process also works in reverse: If the AP has a block of ETF shares it no longer wants (usually because it bought them from other market participants), the AP presents these shares for redemption to the ETF manager and receives in return the basket of underlying securities, which the AP can then sell in the market if it chooses. This basket often has the same security composition as the creation basket, but it may be different if the ETF portfolio manager is trying to sell particular securities for tax, compliance, or investment reasons. The basket of securities the AP receives when it redeems the ETF shares is called the **redemption basket**.

Although the actual process of exchanging baskets and blocks of ETF shares happens after the markets are closed, the AP is able to execute ETF trades throughout the trading day because the AP knows the security composition of the basket needed for ETF share creation or redemption, because of the fund's daily holdings disclosure to APs. If, during the course of the trading day, the AP wants to sell 50,000 shares of an ETF to investors in the secondary market, the AP can do so while simultaneously buying the securities in the creation basket. If the ETF and the securities in the creation basket are fairly priced, the AP faces no economic exposure in this transaction, because the value of the ETF shares sold and the value of the creation basket purchased are identical.

Why would APs engage in these transactions? Because there's a financial incentive to do so. The creation/redemption mechanism is key to keeping the price of an ETF in a tight range around the NAV of the portfolio of securities it holds, and it rewards the AP for this activity.

When the value of the security basket is different from the value of the corresponding ETF shares it represents, a potential arbitrage opportunity exists for APs to step in and transact in the ETF market. If the current per-share market value of the basket of underlying securities is greater than the quoted price of the ETF shares, the AP can simultaneously sell (or short) the basket of securities and buy ETF shares, to make a profit. In this situation, where the ETF share is undervalued, the ETF is said to be trading at a discount. If shares of the ETF are quoted at a higher price than the per-share market value of the basket of securities, the ETF is trading at a premium, and the AP can make a profit by simultaneously selling the ETF shares in the market and buying the basket of securities.

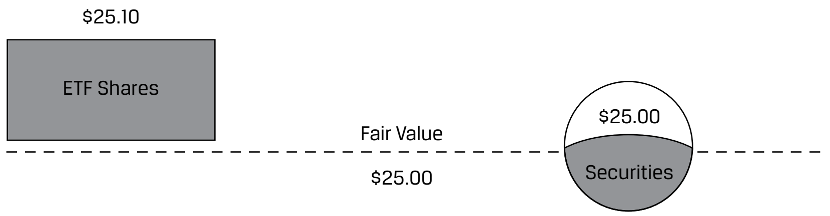
Because prices of the ETF and the basket securities are continuously changing on the basis of market conditions, APs monitor both for discrepancies, looking for opportunities to make arbitrage profits. The factors that drive the width of the ETF's bid–ask spread and trading range around intraday NAV include the cost of arbitrage (buying the securities and selling the ETF) and a risk premium to compensate for volatility and liquidity risk (ongoing volume in the securities and the ETF).

The *arbitrage gap*—the price(s) at which it makes sense for ETF market makers to step in and create or redeem shares—vary with the liquidity of the underlying securities and a variety of related costs; in some ETFs, the gap can be as small as the minimum tick size in the local market (e.g., –$0.01 in the US markets), whereas for other ETFs with underlying securities that are hard to trade (e.g., high-yield bonds), the arbitrage gap can be more than 1% wide. For any ETF, however, the gap creates a band or range around its fair value inside which the ETF will trade. In other words, arbitrage keeps the ETF trading at or near its fair value.

ETF share creation.

Let us examine how this works in practice. In the scenario shown in Exhibit 1, the ETF is trading in the market at $25.10. The fair value of the ETF based on its underlying securities, however, is only $25.00. So, an AP will step in to transact and buy the basket of securities (at ETF fair value of $25.00) and simultaneously sell ETF shares on the open market for $25.10, realizing the $0.10 per share difference. (The AP may choose to create additional ETF shares by exchanging the basket securities for ETF shares with the fund's issuer).

Exhibit 1: An ETF Share Price at a Premium to NAV

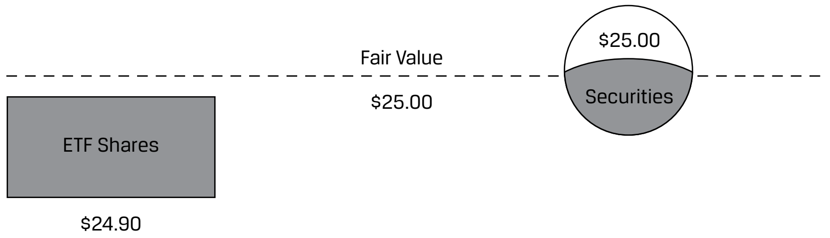


This action puts downward pressure on the ETF price because the AP is selling shares out into the market and puts upward pressure on the prices of the underlying securities because the AP went out into the market and bought the underlying shares. APs will repeat this process until no further arbitrage opportunity exists.

ETF share redemption

As shown in Exhibit 2, the price of the ETF is $24.90. The fair value of the underlying stocks is $25.00. Here, the AP market maker steps in and purchases ETF shares on the open market while simultaneously selling the stocks on the exchange, realizing the $0.10 per share price difference. Once again, if the share price continues to be at a discount, the AP will continue this process until no further arbitrage opportunity exists. (The AP may choose to redeem ETF shares by exchanging them for the basket securities with the fund's issuer).

Exhibit 2: An ETF Share Price at a Discount to NAV



These profit-making scenarios do not include the costs that the APs incur related to ETF trading or any fees the issuer may charge for creating or redeeming shares. The AP generally pays all trading costs associated with buying or selling the securities in the baskets or the ETF shares and pays an additional fee to the ETF provider to cover processing fees associated with creation/redemption activities. APs may also have settlement costs, taxes, or other expenses based on their local markets and the markets for the underlying securities of the ETF.

The scenarios also do not account for risks in trading the basket of securities. If the underlying securities are difficult to access contemporaneously (for instance, if a US-listed ETF holds Japanese securities), then the AP will have to wait before completing one half of the transaction (e.g., selling the ETF shares but waiting until the Japanese market opens to buy the basket securities). These timing differences create uncertainty, which will generally cause the AP to wait for a wider arbitrage gap before stepping in. Similarly, if the basket securities are illiquid (such as high-yield bonds), the AP may need additional time to buy or sell the holdings. In both cases, the AP bears the market risk of the basket transaction.

A significant advantage of the ETF creation/redemption process is that the AP absorbs all costs of transacting the securities for the fund's portfolio. APs pass these costs to investors in the ETF's bid–ask spread, incurred by ETF buyers and sellers. Thus, non-transacting shareholders of an ETF are shielded from the negative impact of transaction costs caused by other investors entering and exiting the fund. In contrast, when investors enter or exit a traditional mutual fund, the mutual fund manager incurs costs to buy or sell investments arising from this activity, which affect all fund shareholders. This makes the ETF structure inherently more fair: Frequent ETF traders bear the cost of their activity, whereas buy-and-hold ETF shareholders are shielded from those costs.

Additionally, because creation and redemption happen in kind, they allow the ETF's portfolio managers to manage the cost basis of their holdings by selecting low-basis holdings for redemptions, leading to greater tax efficiency. Put simply, when an issuer is presented with a redemption request from an AP, the issuer can select which tax lots of the underlying securities to deliver. In addition, issuers may choose to publish customized redemption baskets, which allows them to target specific low-basis securities for removal from the portfolio. By delivering out shares that were originally acquired at low costs, the issuer can continuously raise the average acquired cost (or cost basis) of each position, thereby minimizing the position's unrealized gains.

The ETF issuer has the ability to determine how the process works for a fund. If the issuer requires that a creation basket be 200,000 shares instead of 50,000 shares, the AP will have less incentive to step in to arbitrage when net new demand is lower than 200,000 shares per day. Basket sizes range from 10,000 shares to 600,000 shares. If the ETF holds highly illiquid securities, the issuer can alter the basket that APs must deliver, thereby lowering the costs of creation. In the most extreme case, the fund may allow for the creation of ETF shares in exchange for cash. Issuers can also charge minimal or large fees for creation and redemption, which affect an AP's profit consideration and transactions, to keep prices in line with fair value. Consider the fee of $50 for the Vanguard Short-Term Inflation-Protected Securities ETF (VTIP) versus the fee of $28,000 for the Vanguard FTSE All-World Ex-US Small-Cap ETF (VSS).

CREATION/REDEMPTION ASSET CLASS DIFFERENCES

The creation/redemption mechanism described is broadly representative of how most ETFs work, regardless of their particular legal structure. Depending on the asset class, however, some differences exist.

Fixed-income ETFs generally hold large amounts of bonds, which may be illiquid to trade (for example, a high-yield municipal bond ETF holds securities that might trade only every few days). Because of this, ETF issuers may choose not to do in-kind creations and redemptions but instead accept equivalent cash value. This makes the process easier for APs, encouraging greater ETF activity, but does result in trading costs and tax impact for the ETF. ETF issuers must balance those costs against the benefit of having the AP participate more actively in the market, keeping spreads tight and the price of the ETF close to fair value.

Similarly, many leveraged and inverse ETFs and commodity ETFs may use cash creation/redemption because it makes managing their underlying swap positions easier. Because swaps are generally negotiated OTC transactions, it would be difficult to have APs participate in increasing or decreasing those swap positions.

Trading and Settlement

There is much confusion in the investor community regarding the underlying mechanics of ETF trading and settlement. Whether this confusion relates to shorting, how shares are created/redeemed or settled, or how they trade, ETFs are potentially confusing to many investors. From the perspective of an investor buying on the open market, ETFs go through the same settlement and clearing process as other listed stocks. This section explains that process as it applies in two regions.

US settlement: National Security Clearing Corporation and Depository Trust Company.

In the United States, all trades that have been entered into on a given business day are submitted at the end of the day to the National Security Clearing Corporation (NSCC). As long as both parties of a transaction agree that Party 1 sold to Party 2 *N* shares of XYZ stock, the NSCC becomes the guarantor of that transaction—the entity that ensures all parties are immunized against the financial impact of any operational problems—on the evening of the trade, and the trade is considered "cleared." After this point, the buyer is guaranteed beneficial ownership in the stock (or ETF) as of the time the trade was marked "executed," even if something (e.g., bankruptcy) happens to the seller before the trade is settled.

The Depository Trust Company (DTC), of which the NSCC is a subsidiary, holds the book of accounts—the actual list of security holders and ownership. This information is aggregated at the member firm level, rather than at the individual investor level. For instance, the DTC keeps track of how many shares of Microsoft are currently held by J.P. Morgan or Charles Schwab, but Charles Schwab is responsible for keeping track of which of its customers own how many shares.

After each trade is cleared, the DTC then adds up the total of all trades in a process of continuous net settlement. For example, suppose at the end of a trading day the following is true:

- E*TRADE owes Schwab 1,000 shares of SPY.
- Schwab owes Bank of America Merrill Lynch 1,000 shares of SPY.

Then, from the DTC's perspective, Schwab is "whole": It both is owed and owes 1,000 shares of SPY. To settle the day's transactions, E*TRADE's account will be debited the 1,000 shares of SPY and Bank of America Merrill Lynch will be credited 1,000 shares.

The NSCC has two days to complete this process and have each firm review its records and correct any discrepancies. We refer to this two-day period as T+2 (trade date + 2 days). This T+2 settlement process works for the vast majority of ETF transactions.

Market makers receive special treatment on settlement requirements. Because the role of market makers is to make a continuous market in a given security by standing ready to buy or sell the security on the basis of demand/supply imbalances, they are more likely to end up truly short at the end of a given day. Because of the time required to create or borrow ETF shares, market makers are given up to six days to settle their accounts.

European trading and settlement.

In Europe, the majority of ETF owners are institutional investors. Additionally, the market is fragmented across multiple exchanges, jurisdictions, and clearinghouses. This fragmentation results in the use of many different trading strategies by investors in both the primary and secondary markets for ETFs. Fundamentally, trading works the same as in the United States: An investor purchases shares in the secondary market from a market maker or other counterparty. APs use the creation/redemption mechanism, which helps keep the ETF share price in line with its fair value.

The majority of trading happens in negotiated OTC trades between large institutions, and although those trades are reported, they do not appear as "live" or published bids and asks on the public markets prior to their execution. Most ETFs in Europe are also cross-listed on multiple exchanges and may have different share classes available that vary in their treatment of distributions or currency hedging. The fragmented European settlement process means that trades are cleared to one of 29 central securities depositories (or CSDs). This has no direct impact on investors other than the inherent complexity of such a system, which may result in wider spreads and higher local market trading costs.

3 UNDERSTANDING ETFS

- ☐ describe sources of tracking error for ETFs
- ☐ describe factors affecting ETF bid–ask spreads
- ☐ describe sources of ETF premiums and discounts to NAV
- ☐ describe costs of owning an ETF

Among the most important questions an investor can ask about an ETF is, Does the fund deliver on its promise? The best-managed ETFs charge low and predictable investment costs, closely track the indexes on which they are based, and provide investors with the lowest possible tax exposure for the investment objective. Additionally, these funds provide complete, accurate information in their prospectuses and marketing materials and explain the fund's structure, composition, performance, and risks. To best understand an ETF's ability to meet expectations, its expense ratio, index tracking, tax treatment, and potential costs and risks should be considered.

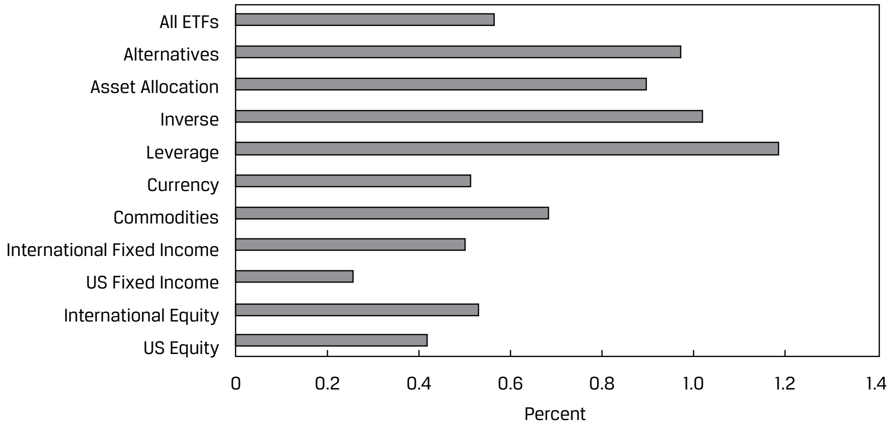
Expense Ratios

Fund expense ratios are often one of the first factors investors look at when evaluating ETFs. ETFs generally charge lower fees than mutual funds, in part because ETF providers do not have to keep track of individual investor accounts, since ETF shares are held by and transacted through brokerage firms. Nor do ETF issuers bear the costs of communicating directly with individual investors. In addition, index-based portfolio management, used by most ETFs, does not require the security and macroeconomic research carried out by active managers, which increases fund operating costs.

The actual costs to manage an ETF vary, depending on portfolio complexity (number of securities held, frequency of rebalancing or strategy implementation, difficulty in maintaining portfolio exposures), issuer size (economies of scale apply), and the competitive landscape.

ETF expense ratios have been one of the most visible areas of competitive differentiation for issuers, which has led to an overall decline in fees. Exhibit 3 shows average US-domiciled ETF expense ratios by asset class at the end of 2018.

Exhibit 3: Average US-Domiciled ETF Expense Ratios by Asset Class at the End of 2018



Sources: ETF.com and FactSet, as of 31 December 2018.

Because the average numbers include complex and expensive funds, they dramatically overstate the cost of accessing the most common ETF investment strategies and indexes.

As of the end of 2018, expense ratios for broad-based, capitalization-weighted indexes were as low as 0.03% for US equities, 0.11% for emerging market equities, and 0.04% for US bonds.

Index Tracking/Tracking Error

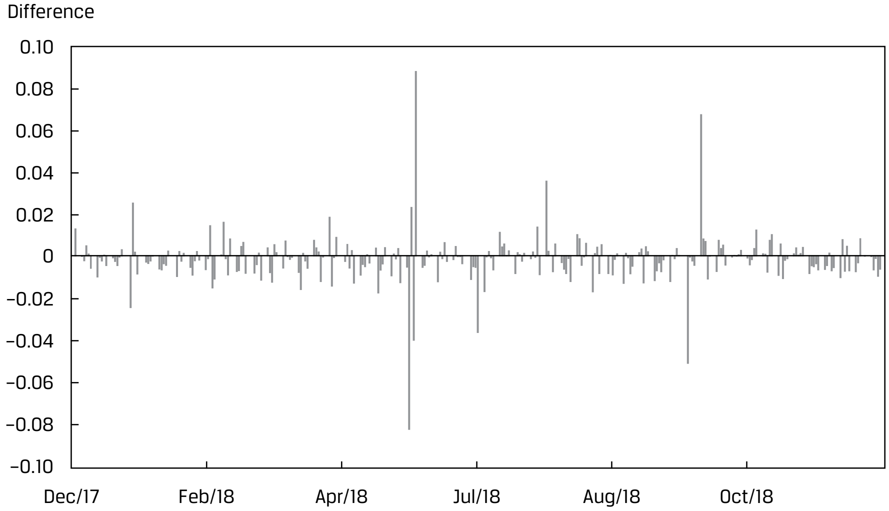
Even though an ETF's expense ratio is useful, it does not fully reflect the cost of holding an ETF. To understand how well an ETF delivers on its mandate, it is critical to assess the ETF's ability to track its underlying index.

For index-tracking ETFs, which represented 98% of the US ETF market as measured by assets under management (AUM) as of December 2018, ETF managers attempt to deliver performance that tracks the fund's benchmark as closely as possible (after subtracting fees). This can be measured by comparing ETF performance with index returns. The comparison can be done using daily or periodic returns but should always include both a central tendency, such as mean or median, and an expression of variability, such as standard deviation or range.

Daily differences.

Index tracking is often evaluated using the one-day difference in returns between the fund, as measured by its NAV, and its index. Exhibit 4 shows the daily tracking difference between the iShares MSCI Emerging Markets ETF (EEM) and its underlying index, the MSCI Emerging Markets Index (EMI), for a one-year period. EMI is a multicurrency international index containing hundreds of illiquid securities in more than 20 emerging markets. The index represents large- and mid-cap stocks in each of these markets. At the end of November 2018, EEM held approximately 900 of the 1,150 constituents in EMI.

Exhibit 4: EEM Daily Tracking Difference Relative to EMI, One-Year Period Ending 30 November 2018



Source: FactSet.

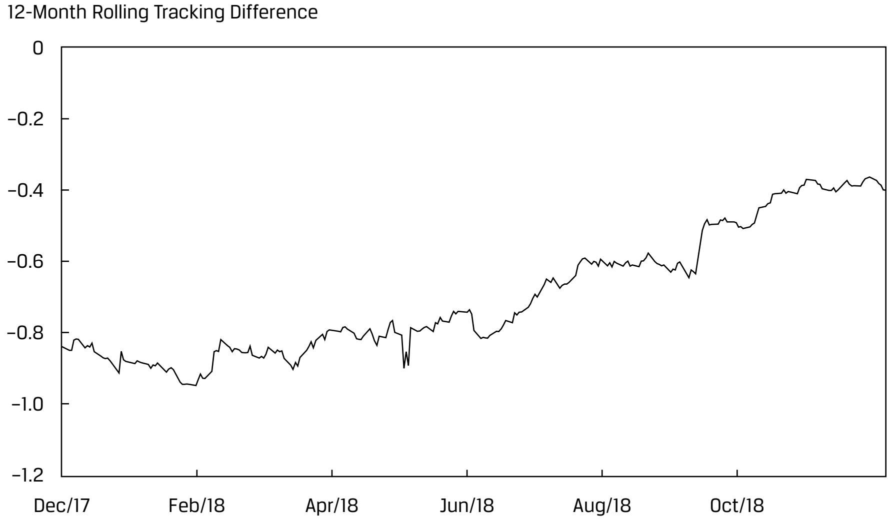
Periodic tracking.

Tracking error is defined as the standard deviation of differences in daily performance between the index and the fund tracking the index, and a reported tracking error number is typically for a 12-month period. Over the period shown, EEM's standard deviation of daily performance differences to its index was 0.012%.

But importantly, tracking error does not reveal the extent to which the fund is under- or overperforming its index or anything about the distribution of errors. Daily tracking error could be concentrated over a few days or more consistently experienced. Therefore, tracking error should be assessed with the mean or median values.

An alternative approach is to look at tracking differences calculated over a longer holding period. A series of rolling holding periods can be used to represent both central tendencies and variability. This approach allows investors to see the cumulative effect of portfolio management and expenses over an extended period. Exhibit 5 shows the 12-month rolling return (or cumulative annual) tracking difference between EEM and its index.

Exhibit 5: EEM 12-Month Rolling Tracking Difference Relative to EMI, One-Year Periods Ending 30 November 2018



Source: FactSet.

One benefit of the rolling annual analysis is that it allows for comparison with other annual metrics, such as the fund's expense ratio. All else equal, one would normally expect an index fund to underperform its benchmark on an annual basis by the amount of its expense ratio. In Exhibit 5, EEM's median tracking difference of 0.79% exceeded its 0.69% expense ratio by 0.10%. Notably, the range of EEM's annual tracking difference showed some variability, with underperformance as low as 0.38% and as high as 0.95%.

Sources of tracking error.

Numerous factors can account for differences between an ETF's expected and actual performance and the range of results with respect to its index. Because of this, funds tracking the same underlying index can have very different index tracking results. Sources of benchmark tracking error include the following:

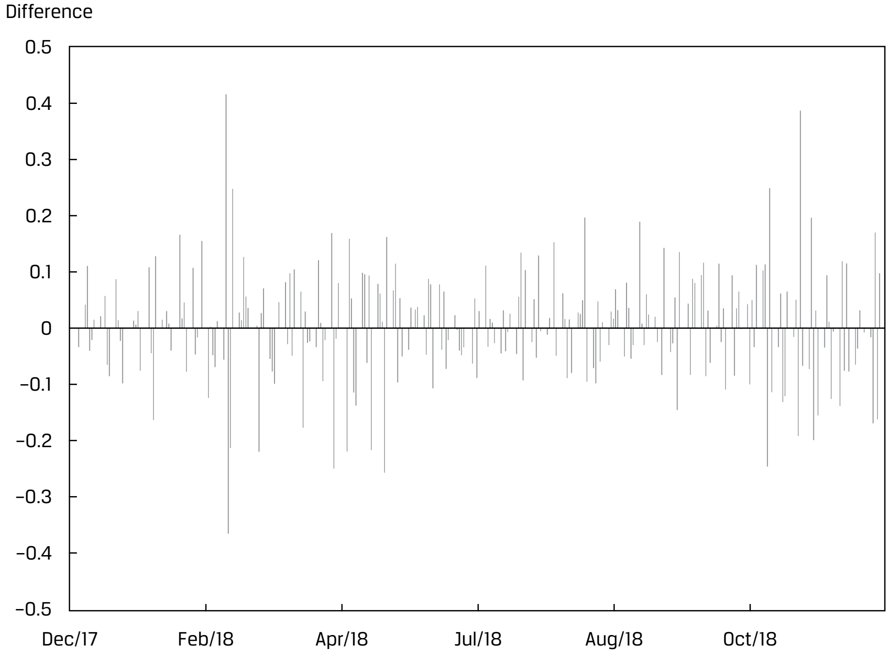
- Fees and expenses—Index calculation generally assumes that trading is frictionless and occurs at the closing price. A fund's operating fees and expenses reduce the fund's return relative to the index.
- Representative sampling/optimization—Rather than fully replicate the index, funds may hold only a subset of index securities to track the benchmark index.
- Depositary receipts and other ETFs—Funds may hold securities that are different from those in the index, such as American depositary receipts (ADRs), global depositary receipts (GDRs), and other ETFs.
- Index changes—Funds may trade index changes at times and prices that are different from those of the benchmark tracked.
- Fund accounting practices—Fund accounting practices may differ from the index calculation methodology—for example, valuation practices for foreign exchange and fixed income.
- Regulatory and tax requirements—Funds may be subject to regulatory and tax requirements that are different from those assumed in index methodology, such as with foreign dividend withholding.

- Asset manager operations—ETF issuers may attempt to offset costs through security lending and foreign dividend recapture. These act as "negative" costs, which enhance fund performance relative to the index.

Fees and expenses.
As outlined in the prior sections, fund operating expenses vary by ETF, but all else equal, one would normally expect an index fund to underperform its benchmark on an annual basis by the amount of its expense ratio.

Representative sampling/optimization.
For funds tracking index exposure to small or illiquid markets, owning every index constituent can be difficult and costly. Therefore, fund managers may choose to optimize their portfolios by holding only a portion, or representative sample, of index securities. A striking example is the SPDR S&P Emerging Asia Pacific ETF (GMF). As of 7 December 2018, GMF held only 763 of the 2,342 securities in the S&P Asia Pacific Emerging BMI Index. As shown in Exhibit 6, sampling has caused some sizable discrepancies between the fund's daily return and the index.

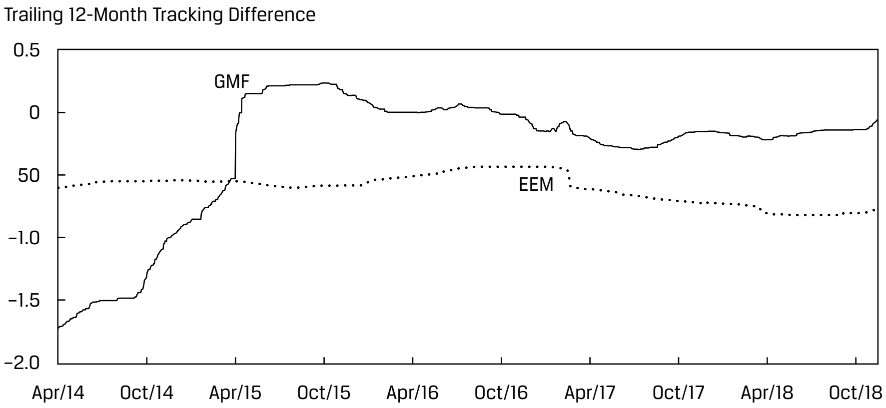
Exhibit 6: GMF Daily Tracking Difference Relative to the S&P Asia Pacific Emerging BMI Index, One-Year Period Ending 30 November 2018



Source: FactSet.

Sampling, or optimization, can affect long-term tracking in two ways. First, it can make the median value unpredictive of future median values, especially if market regimes shift. Second, it dramatically expands the range of results. Exhibit 7 and the table below illustrate these effects, using trailing 12-month (TTM) rolling comparisons. Exhibit 7 contrasts EEM's median trailing 12-month tracking difference with GMF's more variable results.

Exhibit 7: Trailing 12-Month Tracking Difference: EEM and GMF



Source: FactSet.

A high level of optimization causes GMF's portfolio to underperform in certain market regimes and outperform in others. Looking at the differences between GMF and its underlying index explains why. As of 30 November 2018, GMF's median constituent market cap was $2.8 billion, whereas the S&P Asia Pacific Emerging BMI's was $0.695 billion, indicating that by holding approximately one-third of index constituents, GMF's portfolio omits many of the index's mid-caps and small-caps. Therefore, GMF will likely underperform the index during times when emerging market mid-caps and small-caps outperform emerging market large-caps, and vice versa.

As illustrated in the following table, GMF's tracking range—the spread between its maximum and minimum trailing 12-month tracking difference—is nearly 4 times that of EEM. A higher level of optimization within GMF causes it to have a wider range of tracking difference relative to its index.

EEM and GMF Tracking Range, One Year Ending 30 November 2018

	EEM	GMF
Maximum TTM	–0.38%	1.14%
Minimum TTM	–0.95%	–0.81%
Range	0.57%	1.95%

Source: FactSet.

Representative sampling/optimization, therefore, enhances or detracts from fund returns relative to the index depending on whether ETF portfolio holdings outperform or underperform those in the index. Compared with a full replication approach, representative sampling/optimization introduces greater potential for tracking error.

Depositary receipts and ETFs.

When local market shares are illiquid, ETF portfolio managers may choose to hold depositary receipts instead of local constituent shares. Although the economic exposure is equivalent, exchange trading hours for these securities differ. Differences in trading hours and security prices create discrepancies between portfolio and index values. Similarly, ETF issuers may choose to hold ETFs as underlying holdings. This also creates discrepancies between fund NAV and index value, because the ETFs' holdings are valued at their closing market price and not their NAV.

Index changes.

An index provider will periodically change index constituents or weights to comply with its index methodology. In the real world, portfolio managers may transact these changes before or after the effective date or time of the index change/closing prices, at different prices. The more volatile the market, the wider the bid–offer spreads and range of traded prices. ETF portfolio managers can use the creation/redemption process to manage rebalance trades, by cooperating with APs to ensure market-on-close pricing on the rebalance date, thus minimizing this source of tracking error.

Fund accounting practices.

Differences in valuation practices between the fund and its index can create discrepancies that magnify daily tracking differences. Some ETF issuers follow the index industry's convention of establishing (striking) currency valuations using WM/Reuters rates, which are set at 4:00 p.m. GMT (11:00 a.m. EST), whereas others conform to established mutual fund industry practices of striking currency valuations at the close of ETF trading. In the United States, equity markets close at 4:00 p.m. ET. Many fixed-income ETF portfolios value bond positions at the time of the equity market close, in keeping with ETF industry custom. However, fixed-income indexes often follow the bond market's practice of valuing bonds at an earlier time. These practices may create valuation discrepancies between the ETF's NAV and the index value, particularly in volatile segments of the bond market, such as long-dated maturities. Valuation discrepancies can also occur for ETFs holding futures, foreign securities, physical metals, and currencies held in specie.

Regulatory and tax requirements.

Regulatory and tax requirements may cause a fund to mis-track its index. For example, non-domestic holders of a nation's securities owe tax on distributions received from securities of companies domiciled in that nation. The tax withholding rate charged is determined by treaty and investor domicile. Index providers who offer a "net" return series adjust the dividends received to account for the tax charged, usually from the point of view of US-domiciled investors. Index providers may use rates different from those experienced by the ETF, however, which can create return differences between the ETF and its index. For many years, Brazil imposed a tax on foreign investments coming into the country. Although this tax did not affect the closing prices of the local stocks and, therefore, was not reflected in index calculation methodology, non-local ETFs domiciled outside Brazil paid this tax whenever they acquired Brazilian stocks. This caused fund underperformance relative to the index.

Asset manager operations.

ETF issuers may engage in security lending or foreign dividend recapture to generate additional income to offset fund expenses. These can be considered "negative" costs. Many ETFs (and mutual funds) lend a portion of their portfolio holdings to short sellers. In exchange, the ETF receives a fee and earns interest on the collateral posted by the borrower (generally, overnight fixed-income securities), which creates income for the portfolio. Because the index calculation does not account for securities-lending income, it is a source of tracking error. Asset managers may work with foreign governments to minimize tax paid on distributions received.

Tax Issues

Two kinds of tax-based evaluations must be made for all ETFs: First, the investor must consider the likelihood of an ETF distributing capital gains to shareholders. Second, the investor must consider what happens when the investor sells the ETF. These two actions are distinct; the tax efficiency of a fund regarding its capital gains distributions has no relation to its tax efficiency at the time of investor sale.

Capital Gains Distributions

The issue of capital gains distributions affects all investors in taxable accounts. In general, funds must distribute any capital gains realized during the year. Funds typically make these distributions at year-end, although they may make them quarterly or on another periodic schedule.

ETFs are said to be "tax fair" and "tax efficient" because they have certain advantages over traditional mutual funds regarding capital gains distributions. On average, they distribute less in capital gains than competing mutual funds for two primary reasons.

Tax fairness.
In a traditional mutual fund, when an investor sells, the fund must (with a few exceptions) sell portfolio securities to raise cash to pay the investor. Any securities sold at a profit incur a capital gains charge, which is distributed to remaining shareholders. Put another way, in a traditional mutual fund, shareholders may have to pay tax liabilities triggered by other shareholders redeeming out of the fund.

In contrast, an investor sells ETF shares to another investor in the secondary market. The ETF manager typically does not know that the sale is occurring and does not need to alter the portfolio to accommodate this transaction. Thus, the selling activities of individual investors in the secondary market do not require the fund to trade out of its underlying positions. If an AP redeems ETF shares, this redemption occurs in kind. In markets where redemptions in kind are allowed, this is not a taxable event. Thus, redemptions do not trigger capital gain realizations. This aspect is why ETFs are considered "tax fair": The actions of investors selling shares of the fund do not influence the tax liabilities for remaining fund shareholders.

Tax efficiency.
The redemption process allows portfolio managers to manage the fund's tax liability. When an authorized participant submits shares of an ETF for redemption, the ETF manager can choose which underlying share lots to deliver in the redemption basket. By choosing shares with the largest unrealized capital gains—that is, those acquired at the lowest cost basis—ETF managers can use the in-kind redemption process to reduce potential capital gains in the fund. Tax lot management allows portfolio managers to limit the unrealized gains in a portfolio.

Other Distributions

Other events, such as security dividend distributions, can trigger tax liabilities for investors but the treatment varies by region, so investors must ensure they understand the tax treatment specific to each fund's domicile, legal structure, and portfolio type.

Taxes on Sale

In most jurisdictions, ETFs are taxed according to their underlying holdings. For example, in the United States, an ETF holding equities or bonds will itself be subject to the same capital gain, dividend, and return-of-capital tax rules that apply to its underlying stock or bond holdings. There can be nuances in individual tax jurisdictions, however, that require investor analysis. For example, in the United States, exchange-traded notes tracking commodity indexes are treated differently from exchange-traded funds holding commodity futures contracts, creating a preferential tax treatment. A thorough analysis of ETF efficiency should take into account the ETF structure, the local market's taxation regime, and the individual tax situation of the end investor.

ETF Trading Costs

In comparing ETF and mutual fund costs, the usual starting point is management fees, which are often lower for an ETF because most are index based and traded in a highly competitive market. Other important costs should be considered, however.

An ETF has the advantage that it can be purchased whenever exchanges are open—as well as at closing NAV of the fund (similar to mutual fund purchases and sales) when a transaction is large enough to qualify for a creation or redemption. ETF investors usually pay a commission and incur a trading cost related to the liquidity factors associated with the ETF. The trading, or market impact, costs are influenced by the bid–ask spread of the ETF, the size of the trade relative to the normal trading activity of the ETF, and the ease of hedging the ETF by the market-making community. The closing price of the ETF on the exchange may include a premium or discount to the NAV, driven by supply and demand factors on the exchange and the market impact costs of executing an exchange transaction.

ETF Bid–Ask Spreads

One of the most important drivers of ETF bid–ask spreads and liquidity is the market structure and liquidity of the underlying securities held. Fixed-income securities, which trade in a dealer market, tend to have much wider bid–ask spreads than large-capitalization stocks. The bid–ask spread of an ETF holding stocks traded in other markets and time zones is influenced by whether the markets for the underlying stocks are open during the hours in which the ETF trades. For specialized ETFs—such as those tracking commodities, volatility futures, or even small-cap stocks—bid–ask spreads can be wide simply because the risk of holding a position even for a short period of time can be high. For some ETFs, even though the underlying securities are liquid, bid–ask spreads may be wide simply because the ETF trades so infrequently the market maker or liquidity provider may need to carry ETF positions for some time before they accumulate sufficient size to create or redeem. Generally, as long as the liquidity in the underlying securities is adequate or hedging instruments can be easily sourced, an ETF trade can usually be executed in a cost-effective manner.

The primary factors that determine the width of the quoted bid–ask spread for a particular transaction size are the amount of ongoing order flow in the ETF, as measured by daily share volume ; the amount of competition among market makers for that ETF ; and the actual costs and risks for the liquidity provider. The bid–ask spread represents the market maker's price for taking the other side of the ETF transaction, which includes the costs and risks to carry the position on his books or to hedge the position using underlying securities or closely related ETFs or derivatives.

More specifically, ETF bid–ask spreads are generally less than or equal to the combination of the following:

- ± Creation/redemption fees and other direct trading costs, such as brokerage and exchange fees
- + Bid–ask spreads of the underlying securities held in the ETF
- + Compensation (to market maker or liquidity provider) for the risk of hedging or carrying positions for the remainder of the trading day
- + Market maker's desired profit spread, subject to competitive forces
- − Discount related to the likelihood of receiving an offsetting ETF order in a short time frame

Large, actively traded ETFs have narrow bid–offer spreads and the capacity (or liquidity) for large transaction sizes. For very liquid US-listed ETFs, such as SPY (the SPDR S&P 500 ETF), or EEM (the iShares MSCI Emerging Markets ETF), buyers and sellers are active throughout the trading day and market makers have a high likelihood of finding the other side or hedging larger orders. Therefore, because most of these ETF trades are matched quickly and never involve the creation/redemption process, the first three factors do not contribute heavily in their spreads. For liquid ETFs, the bid–ask spread can be significantly tighter than the spreads on the underlying securities.

The quoted ETF bid–ask spread, however, is generally for a specific, usually small, trade size and does not always reflect ETF liquidity for larger transactions (more than 10% of average daily volume). Larger trades may best be handled by negotiation, involving work with capital market specialists at ETF managers and broker/dealer ETF desks to understand the various ETF execution options and associated trading costs.

Exhibit 8 shows the asset-weighted average and median bid–ask spreads for various ETF categories traded in the United States.

Exhibit 8: Average and Median Bid–Ask Spreads for US-Traded ETFs

US-Traded ETF Category	AUM ($ millions)	Average Spread ($ asset-weighted)	Median Spread
US Equity	1,871,942	0.03%	0.16%
International Equity	731,251	0.05%	0.24%
US Fixed Income	589,851	0.02%	0.14%
International Fixed Income	65,159	0.06%	0.24%
Commodities	62,620	0.05%	0.24%
Leveraged	29,633	0.29%	0.32%
Inverse	11,315	0.10%	0.21%
Asset Allocation	9,318	0.21%	0.29%
Alternatives	4,388	0.18%	0.38%
All US-Traded ETFs*	3,377,276	0.04%	0.20%

** Includes currency ETFs in addition to ETFs listed. Total currency ETF assets are $1,799 million.*
Source: FactSet, as of the end of December 2018, based on 60-day averages.

US equity and fixed-income ETFs have the tightest asset-weighted spreads. International equity and international fixed-income spreads are wider, because the underlying securities trade in different market structures, making it difficult to price simultaneously, and because the underlying security exchanges may be closed during a portion of the US trading day. ETF categories representing longer-term strategies, such as asset allocation and alternatives, are less actively traded and have lower asset levels and wider spreads, in part because they have less ongoing two-way order flow and, therefore, depend more on market makers to source liquidity through the underlying securities. Bid–ask spreads are dynamic, vary by trade, and tend to widen when market volatility increases or when significant information relating to the underlying index securities is expected.

Understanding spreads for non-equity ETFs is more complex. Although the fixed-income ETFs give investors access to a portfolio of debt securities trading with transparent bid–ask spreads in the stock market (via the ETF), the actual market for the underlying bonds is far less transparent with OTC trading, in which traders at banks and large bond desks offering quotes on demand without posting bids or offers on an exchange.

Unlike actively traded US Treasury securities, both corporate debt and high-yield bonds, as well as some municipals and international bonds, trade actively only around the time of issuance, after which they may be held until maturity. Therefore, bond ETFs that track indexes containing corporate and high-yield debt often invest only in a subset of the most liquid high-yield securities. Their bid–ask spreads tend to be wider than those of ETFs based on stocks or US Treasuries because of the risk to dealers in hedging inventory and the default risk of the securities, especially in periods of weak economic conditions.

Premiums and Discounts

In addition to commissions and bid–ask spreads, ETF premiums and discounts are also important components of ETF trading costs.

At the end of the trading day, each ETF has an end-of-day NAV at which shares can be created or redeemed and with which the ETF's closing price can be compared. Most investors rely on return calculations based on this closing NAV. NAV is intended to be an accurate assessment of the ETF's fair value. This is the case when the underlying securities trade on the same exchange as the one where the ETF is listed (or trades), because these securities trade in the same market structure and have the same closing price time as the ETF.

During the trading day, exchanges disseminate ETF **iNAVs**, or "indicated" NAVs; iNAVs are intraday "fair value" estimates of an ETF share based on its creation basket composition for that day. An ETF is said to be trading at a premium when its share price is higher than iNAV and at a discount if its price is lower than iNAV.

The calculation for end-of-day and intraday premiums/discounts is as follows:

End-of-day ETF premium or discount (%)

= (ETF price − NAV per share)/NAV per share.

Intraday ETF premium or discount (%)

= (ETF price − iNAV per share)/iNAV per share.

Like tracking error, premiums/discounts are driven by a number of factors, including timing differences and stale pricing.

Timing differences.

NAV is often a poor fair value indicator for ETFs that hold foreign securities because of differences in exchange closing times between the underlying (e.g., foreign stocks, bonds, or commodities) and the exchange where the ETF trades. For example, if a commodity held in the fund stops trading in the futures market at 3:00 p.m., the issuer may elect to retain that price for a 4:00 p.m. valuation. If a fund holds securities in a different currency, it may choose to "strike" or value the currency at 4:00 p.m. ET—or occasionally, at 4:00 p.m. London time. In the case where international stocks are held in US-traded ETFs, the NAV may be based on a market closing price in Asia or Europe that occurred hours ahead of when the ETF stops trading on the US exchange.

Because bonds do not trade on an exchange, no true "closing prices" are available for valuing the bonds in a portfolio. Instead, ETF issuers rely on bids from bond desks or pricing services for proxy prices. In the case of bonds that have not traded near the close of the dealer market, index providers and bondholders typically use pricing services for bond valuation. These pricing services often use more liquid bonds that have similar features to estimate where the non-traded bond would have closed.

Sometimes, bond pricing model inputs reflect the price at which a dealer is willing to buy the bonds and the risk and cost to a dealer in carrying the bonds in inventory. In such cases, the ETF's closing price is often higher than the bid prices of the underlying bond holdings used to calculate NAV, making it appear that the ETF is at a premium. During times of market stress, few bonds may trade, leaving pricing services without updated inputs for their models. Like ETFs holding foreign securities, this causes NAVs to be "stale" and, in this case, with possibly too high a valuation given market conditions. In this case, fixed-income ETFs with sufficient trading volume may appear to be trading at discounts to NAV. In these cases, by reflecting the market's most current assessment of value, liquid ETFs become "price discovery" vehicles.

ETFs also provide price discovery for after-hours markets. For example, US-listed ETFs holding European stocks trade until 4:00 p.m. ET, hours after European markets have closed. In these cases, premiums or discounts resulting from closed underlying markets are not mispricing; rather, they are the market's best estimate as to where the fund holdings would trade if the underlying markets were open.

Stale pricing.

ETFs that trade infrequently may also have large premiums or discounts to NAV. If the ETF has not traded in the hours leading up to the market close, NAV may have significantly risen or fallen during that time owing to market movement. In this case, comparing the last ETF trade price—for example, at 1:00 p.m.—with the end-of-day 4:00 p.m. NAV would result in a premium (or discount) if the market and corresponding NAV fell (or rose) sharply between 1:00 and 4:00 p.m.

This situation can be compounded if days or weeks elapse between the ETF's trades. Some premium/discount calculations use a strict last price input, whereas others use a closing midpoint. The strict pricing will quote the last trade price, no matter how distant the ETF trade date, which can lead to severe premiums or discounts because NAVs are updated on the basis of the latest market closing prices while the ETF price remains unchanged at last trade.

EXAMPLE 1

Comparison of US ETF Trading Costs

A good way to assess the liquidity and potential trading costs of ETFs is to compare various measures of trading activity among similar funds. Exhibit 9 shows trading measures for some of the most liquid ETFs—the SPDR S&P 500 ETF (SPY), the iShares Core S&P 500 ETF (IVV), and the Vanguard S&P 500 ETF (VOO) benchmarked to the S&P 500 Index; another large-cap ETF, the iShares MSCI USA Equal Weighted ETF (EUSA), benchmarked to the MSCI USA Equal Weighted Index; and a liquid small-cap ETF, the iShares Russell 2000 ETF (IWM), benchmarked to the Russell 2000 Index.

Exhibit 9: Selected US Equity Index ETF Trading Measure Comparison

ETF Ticker	SPY	IVV	VOO	EUSA	IWM
Benchmark Index	**S&P 500**	**S&P 500**	**S&P 500**	**MSCI USA Equal Weighted**	**Russell 2000**
Volume in US dollars					
Daily average volume	24.47 billion	1.22 billion	819.28 million	1.32 million	3.90 billion
Median volume	20.23 billion	1.08 billion	739.69 million	0.94 million	3.81 billion
Other trading characteristics					
Average spread (%)	0.00%	0.01%	0.01%	0.12%	0.01%
Average spread ($)	$0.01	$0.03	$0.03	$0.07	$0.01
Median premium/discount (%)[a]	0.00%	0.00%	0.00%	0.04%	0.01%

ETF Ticker	SPY	IVV	VOO	EUSA	IWM
Benchmark Index	S&P 500	S&P 500	S&P 500	MSCI USA Equal Weighted	Russell 2000
Maximum premium (%)[a]	0.12%	0.13%	0.18%	0.96%	0.12%
Maximum discount (%)[a]	–0.19%	–0.11%	–0.08%	–0.38%	–0.13%

[a] *Over previous 12 months.*

Source: FactSet, as of 7 November 2018.

SPY, the largest ETF by AUM and the first ETF traded in the United States, is one of the most liquid securities in the world. IVV and VOO, with the same benchmark, are used more by intermediate- and longer-horizon investors but also have very tight spreads because of liquidity in the underlying securities and ease of hedging for market makers. SPY trades a median of $20 billion a day, compared with a median of $1 billion for IVV. The average bid–ask spread shows that both are highly liquid. In addition, both have tight premiums and discounts to NAV.

In contrast, EUSA has a larger spread, 0.12%. The lower liquidity and higher trading cost for EUSA can be attributed to the fact that the benchmark index does not have futures and other index products available for hedging use by market makers. The MSCI USA Equal Weighted Index also includes close to 600 stocks—100 more than the S&P 500 Index has.

IWM, benchmarked to the Russell 2000 Index of US small-cap stocks, holds far more securities than any of the previously mentioned ETFs, and many are small-cap stocks that have wide spreads. IWM, however, trades with spreads and premiums/discounts close to those of SPY.

How is that possible? First, trading activity in IWM is high (median daily dollar volume of $4 billion) and continuous throughout the trading day. Second, the Russell 2000 Index has an active futures market, making it easy for market makers and APs to quickly hedge the risk of large trades.

Exhibit 10 shows three US fixed-income ETFs—one US-Treasury based and two benchmarked to US high-Yield indexes. All three are among the most liquid fixed-income ETFs and have tight average bid–ask spreads. The iShares iBoxx $ High Yield Corporate Bond ETF (HYG) is the most liquid, with median daily volume of $1.4 billion and a higher median premium (0.20%) than the iShares 20+ Year Treasury Bond ETF (TLT). These positive median premiums indicate that the SPDR Bloomberg Barclays High Yield Bond ETF (JNK) and HYG have been in a net demand position over most of the 12-month period covered in Exhibit 10 and investors have typically paid above fair value for ETF access to a high-yield portfolio.

The maximum premium and discount have generally been much larger for bond ETFs compared with the equity ETFs shown in Exhibit 9. This is because the underlying fixed-income securities trade in a dealer market and are not continuously priced. In this case, the fixed-income ETFs, which trade on an exchange with more continuous pricing, may be a better reflection of true supply and demand for the portfolio because the underlying bonds may not trade as frequently, particularly in extreme market conditions.

Exhibit 10: Selected US Fixed-Income ETF Trading Measure Comparison

ETF Ticker	TLT	JNK	HYG
Benchmark Index	ICE US Treasury 20+ Year Bond Index	Bloomberg Barclays High Yield Very Liquid Index	Markit iBoxx USD Liquid High Yield Index
Volume in US dollars			
Daily average volume	1.04 billion	0.46 billion	1.50 billion
Median volume	0.97 billion	0.41 billion	1.44 billion
Other trading characteristics			
Average spread (%)	0.01%	0.03%	0.01%
Average spread ($)	$0.01	$0.01	$0.01
Median premium/discount (%)[a]	0.03%	0.10%	0.20%
Maximum premium (%)[a]	0.68%	0.41%	0.59%
Maximum discount (%)[a]	−0.52%	−0.67%	−0.75%

[a] *Over the previous 12 months.*

Source: FactSet, as of 7 November 2018.

Total Costs of ETF Ownership

Exhibit 11 provides a summary of cost factors when considering ETFs and mutual funds. Some of these costs are explicit, whereas others are implicit and reflected in net investment returns. Both ETFs and mutual funds typically pay lower institutional commission rates for trades because of their asset size. ETF transaction costs are incurred at purchase and sale regardless of holding period, whereas other costs, such as management fees, increase as the holding period lengthens. Ongoing costs, such as management fees, portfolio turnover, and security lending proceeds, have a consistent impact on investment returns based on holding period. ETF trading costs, such as commissions and bid–ask spreads, are incurred only at purchase and sale, and their return impact diminishes over longer holding periods, whereas management fees and other ongoing costs become a more significant proportion of total costs. Tracking error can be considered a positive or negative implicit cost.

For active short-term ETF investors who trade frequently, the cost of entering and exiting their ETF positions (commissions, bid–ask spreads, premiums/discounts) is a far more significant consideration than management fees, tracking error, and other costs that accumulate over longer holding periods.

ETFs may trade at market prices higher (premiums) or lower (discounts) than NAV, which is based on closing prices for the fund's underlying securities. Premiums and discounts may reflect a lag in the timing of the underlying security valuations relative to current market conditions and can be considered positive costs (in the case of premiums) or negative costs (in the case of discounts).

There are additional implicit trading costs of fund management, such as portfolio turnover costs that are reflected in fund returns. These are incurred within the fund as the portfolio manager buys and sells securities to execute the investment strategy and manage fund cash flows. Portfolio turnover costs reduce returns and affect performance for all investors in the fund. Many ETFs are based on indexes that have lower portfolio turnover than actively managed funds. Taxable gains incurred upon sale can be considered positive costs for the investor, whereas taxable losses represent negative costs. Security lending income for the fund represents negative costs.

Exhibit 11: Cost Factor Comparison—ETFs and Mutual Funds

Fund Cost Factor	Function of Holding Period?	Explicit/Implicit	ETFs	Mutual Funds
Management fee	Y	E	X (often less)	X
Tracking error	Y	I	X (often less than comparable index mutual funds)	(index funds only)
Commissions	N	E	X (some free)	
Bid–ask spread	N	I	X	
Premium/discount to NAV	N	I	X	
Portfolio turnover (from investor flows and fund management)	Y	I	X (often less)	X
Taxable gains/losses to investors	Y	E	X (often less)	X
Security lending	Y	I	X (often more)	X

Trading costs vs. management fees.

To illustrate the effect of management fees versus trading costs, consider an investor who pays a commission of \$10 on a \$20,000 trade (0.05% each way) combined with a 0.15% bid–ask spread on purchase and sale. The round-trip trading cost is, therefore, 0.25% and is calculated as follows:

Round-trip trading cost (%)

$$= (\text{One-way commission \%} \times 2) + (\tfrac{1}{2}\ \text{Bid–ask spread \%} \times 2)$$

$$= (0.05\% \times 2) + (\tfrac{1}{2} \times 0.15\% \times 2)$$

$$= 0.10\% + 0.15\%$$

$$= 0.25\%.$$

For a round-trip trade that happens over a year, 0.25% can be larger than the annual expense ratios of many ETFs. If held for less than a year, the trading costs may be far larger than the expense ratio paid on the ETF.

To see the impact of holding period, consider the 3-month versus 12-month versus 3-year holding period costs for an ETF with a 0.15% annual fee, one-way commissions of 0.05%, and a bid–ask spread of 0.15%. Holding period costs can be calculated as follows:

Holding period cost (%)

= Round-trip trade cost (%) + Management fee for period (%).

Specific holding period costs can be calculated as follows:

$$\text{3-month holding period cost (\%)} = 0.25\% + 3/12 \times 0.15\%$$

$$= 0.29\%.$$

$$\text{12-month holding period cost (\%)} = 0.25\% + 12/12 \times 0.15\%$$

$$= 0.40\%.$$

$$\text{3-year holding period cost (\%)} = 0.25\% + 36/12 \times 0.15\%$$

$$= 0.70\%.$$

Exhibit 12 illustrates that for holding periods of 3 and 12 months, trading costs represent the largest proportion of annual holding costs (0.86% and 0.625%, respectively). Excluding the compounding effect, for a three-year holding period, management fees represent a much larger proportion of holding costs (0.64%).

Exhibit 12: ETF Management Fee and Trading Cost Comparison

Holding Period:	3 Months	12 Months	3 Years
Commission	0.10%	0.10%	0.10%
Bid–ask spread	0.15%	0.15%	0.15%
Management fee	0.0375%	0.15%	0.45%
Total	**0.29%**	**0.40%**	**0.70%**
Trading costs (% of total)	0.86%	0.625%	0.36%
Management fees (% of total)	0.14%	0.375%	0.64%

For broad-based, capitalization-weighted equity index ETFs that have the lowest fees, trading costs represent the largest cost in using an ETF. The longer an ETF is held, the greater the proportion of total costs represented by the management fee component.

Tactical traders will generally choose an ETF on the basis of its liquidity and trading costs (e.g., commissions, bid–ask spreads). In many cases, shorter-term tactical traders may use an ETF with a higher management fee but a tighter bid–ask spread and more active or continuous two-way trading flow to avoid incurring the capital commitment cost of a market maker or the cost of arbitrage for the ETF versus the underlying securities. The size of the management fee is typically a more significant consideration for longer-term buy-and-hold investors.

ETF RISKS

4

☐ describe types of ETF risk

ETFs introduce several unique risks because of their structure, fund holdings, and underlying exposure.

Counterparty Risk

Some ETP (exchange-traded product) legal structures involve dependence on a counterparty. A counterparty failure can put the investor's principal at risk of default or affect a portion of the assets via settlement risk. Likewise, counterparty activity can affect a fund's economic exposure. Therefore, investors should carefully assess counterparty risk.

Although exchange-traded notes (ETNs) trade on exchanges and have a creation/redemption mechanism, they are not truly funds because they do not hold underlying securities. ETNs are unsecured debt obligations of the institution that issues them and are structured as a promise to pay a pattern of returns based on the return of the stated index minus fund expenses. The issuer of the note takes responsibility for setting up the counterbalancing hedges it believes necessary to meet the obligations.

In the United States, ETNs are registered under the Securities Act of 1933 because they are general obligation debt securities of a bank and are not managed by an investment firm for a fee. Similar ETN structures exist in most markets where ETFs are listed.

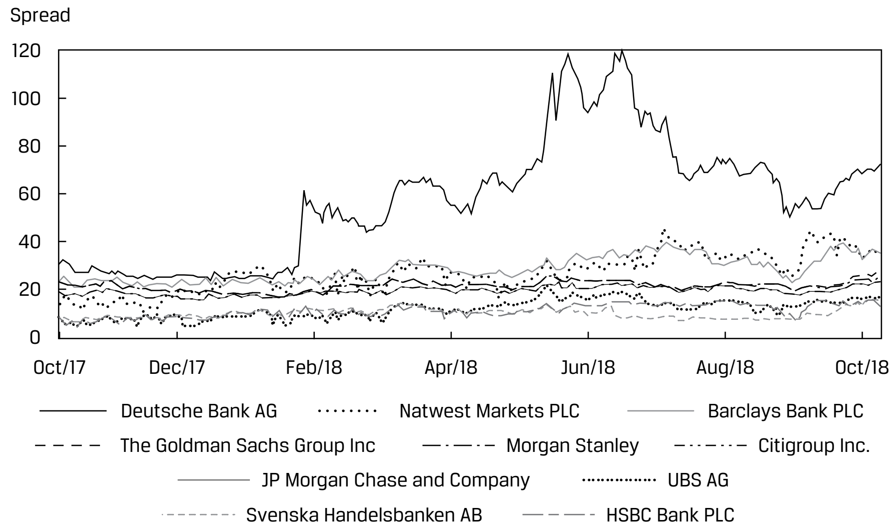
ETNs have the largest potential counterparty risk of all exchange-traded products because they are unsecured, unsubordinated debt notes and, therefore, are subject to default by the ETN issuer. Theoretically, an ETN's counterparty risk is 100% in the event of an instantaneous default by the underwriting bank, and should an issuing bank declare bankruptcy, any ETNs issued by the bank would effectively be worthless. Because baskets of notes may be redeemed back to the issuer at NAV, however, it is likely that only an extremely rapid and catastrophic failure would take investors by surprise. This happened once, in 2008, with three Lehman Brothers–backed ETNs, but it has not happened since.

In the United States, some funds offering exposure to non-US-dollar currencies achieve this via offshore bank deposits. These funds bear default risk at the deposit-holding bank.

Because ETNs and deposit-based ETFs are backed by banks, their default risk can be monitored via the issuing bank's credit default swap (CDS) pricing.

The credit spreads for one-year CDSs by issuer at the end of October 2018 are shown in Exhibit 13.

Exhibit 13: One-Year CDS Spreads for ETN Issuers, 24 October 2018



Source: Bloomberg

The quoted CDS rates represent the cost to insure debt, in basis points per year; so, for example, investors could "insure" $1 million in Goldman Sachs bonds for just under $30,000 per year. Although the insurance rate should never be considered an estimate of actual default risk for a 12-month period, it does provide a reasonable gauge of the relative risk of the various issuers. In general, a one-year CDS rate above 5% should raise significant concerns among investors because it foretells a significant default risk in the year to come.

Settlement risk.

A fund that uses OTC derivatives, such as swaps, to gain market exposure has settlement risk; that is, mark-to-market (unrealized) gains are subject to counterparty default. Such ETFs include many European swap-based funds (or synthetic ETFs), funds using leverage (or geared funds), some currency funds, and some actively managed portfolios. To minimize settlement risk, OTC contracts are typically settled frequently—usually on a daily or weekly basis. This frequent settlement reduces the exposure the swap partners face if a company goes bankrupt, but there is a theoretical risk of counterparty default between settlement periods. In addition, the majority of the contract collateral is held in low-risk instruments, such as US T-bills, at a custodian bank.

Swap exposures are not unique to ETFs. Many mutual funds also use swaps and other derivatives to gain exposure. With ETFs, swap exposures are somewhat transparent because these holdings are disclosed daily by the ETF provider, although full information on counterparties and terms may not be disclosed.

Security lending.

ETF issuers (in addition to traditional mutual fund managers and institutions) lend their underlying securities to short sellers, earning additional income for the fund's investors. Securities lent are generally overcollateralized, to 102% (domestic) or 105% (international), so that the risk from counterparty default is low. Cash collateral is usually reinvested into extremely short-term fixed-income securities with minimal associated risk. At the time of writing, there has been no instance of shareholder loss resulting from security lending in an ETF since ETF product inception in the early 1990s. A well-run security lending program can generate significant income for the ETF issuer, sometimes entirely offsetting the fund's operating expenses. Most ETF issuers credit all profits from this activity back to shareholders, although information about issuer lending programs is sometimes not well disclosed.

Fund Closures

Similar to mutual fund closures, ETF issuers may decide to close an ETF. In such a case, the fund generally sells its underlying positions and returns cash to investors. This activity can trigger capital gain events for investors and the need to find a replacement investment. Primary reasons for a fund to close include regulation, competition, and corporate activity. "Soft" closures—which do not involve an actual fund closing—include creation halts and changes in investment strategy.

Regulations.

Security regulators can change the regulations governing certain types of funds, resulting in forced closure of those funds. For example, commodity futures are under constant regulatory scrutiny, and position limits can make it impossible for some funds to function. In 2018, the Israeli security regulator banned the ETN structure, forcing over 700 products to close and reopen as traditional ETFs.

Competition.

Investors have benefited from a growing number of ETFs and increased competition. As ETFs proliferate, some funds fail to attract sufficient assets and are shut down by the ETF issuer. A fund's assets under management, in addition to those of any competitor, and the ETF's average daily liquidity are indications of market support. Low AUM and trading volumes over a significant period could indicate potential fund closure.

Corporate actions.

Mergers and acquisitions between ETF providers can prompt fund closures. When ETF families merge or are sold to other ETF providers, new ETF owners may close underperforming ETFs (from an asset-gathering perspective) and invest in new, higher-growth opportunities.

Creation and redemption halts.

ETN issuers may halt creations and redemptions. An example of this scenario is when an ETN issuer no longer wants to add debt to its balance sheet related to the index on which the ETN is based. This situation occurred in September 2018, when ETN issuer UBS issued a "sales halt" for its ETRACS Monthly Pay 2xLeveraged Mortgage REIT ETN (MORL), effectively suspending further sales from its outstanding inventory of the ETN and preventing new shares from being created. When creations are halted, the ETN can trade at a substantial premium over fair value, as the arbitrage mechanism breaks down. In this case, MORL traded at a premium of more than 5%. Although all ETFs can theoretically close creations in extraordinary situations, in practice, it happens more commonly with ETNs.

Change in investment strategy.

Some ETF issuers find it easier to repurpose a low-asset ETF from their existing lineup than to close one fund and open another. Issuers simply announce a change in the fund's underlying index—a common occurrence in the ETF industry. Although most index changes result in small adjustments to an ETF's portfolio and economic exposure, these "soft closures" can sometimes result in a complete overhaul, changing exposures to countries, industries, or even asset classes.

Investor-Related Risk

ETFs provide access to sometimes complex asset classes and strategies. For all ETFs, it is important that investors understand the underlying exposure provided by the ETF; otherwise, ETFs may introduce risks to investors who do not fully understand them. For many investors, leveraged and inverse ETFs fall into this category by failing to meet investor expectations. Index methodology (e.g., constituent universe, weighting approach) and the fund's portfolio construction approach are central to understanding an ETF's underlying exposure and related performance.

Leveraged and inverse funds generally offer levered (or geared), inverse, or levered and inverse exposure to a given index and have a daily performance objective that is a multiple of index returns. These products must reset or adjust their exposure daily to deliver the target return multiple each day.

For example, consider a fund offering 300% exposure (3 times, or 3×) to the FTSE 100 Index with a net asset value of £100. It uses swaps to obtain a notional exposure of £300. If the one-day FTSE 100 Index return is 5%, the £300 in exposure becomes £315 (a 5% increase), and the ETF's end-of-day NAV is £115: $100 \times (1 + 3 \times 5\%)$.

In order to deliver 300% of the index's daily performance for the following day, the ETF, now valued at £115, requires notional exposure of £345 for 3 times exposure. Because at the end of the day the ETF has only £315 in exposure, it must reset its exposure—in this case, increasing notional swap exposure by £30.

Exhibit 14 outlines this example.

Exhibit 14: Example of Levered 3× ETF Exposure

	Index Level	One-Day Index Return (%)	3× ETF NAV (£)	Notional Swap Exposure (£)	3× Swap Exposure (£)	Swap Exposure Adjustment (£)
Day 1	100	—	100	300	300	0
Day 2	105	5%	115	315	345	30

If these ETFs are held for longer than a one-day period, the math of compounding and resetting exposure is such that an investor will not see the return multiple—for example, a 200% or −100% return in the case of a 2× ETF or inverse ETF, respectively—over her holding period.

Exhibit 15 presents a levered, inverse fund offering 2 times (−200%) exposure to the S&P 500 Index. The fund (−2× ETF) has a starting net asset value of $100 and uses swaps to obtain notional exposure.

Exhibit 15: Example of Levered and Inverse 2× ETF Daily Return vs. Holding Period Return

	Index Level	One-Day Index Return (%)	Index Period Return (%)	−2× ETF NAV	One-Day ETF Return (%)	2× ETF Holding Period Return (%)
Day 1	100	—		100	—	—
Day 2	110	10%	10%	80	−20%	−20%
Day 3	99	−10%	−1%	96	20%	−4%

Day 1: Both the index and the −2× fund are at a starting level of 100.

Day 2: The index increases to 110, a one-day return of 10%.

The −2× ETF daily return is calculated as follows:

$= -2 \times [(110 - 100)/100]$

$= -2 \times (10\%)$

$= -20\%.$

The −2× ETF NAV is calculated as follows:

$= 100 \times (1 + -0.2)$

$= 80.$

Day 3: The index falls to 99, a one-day return of −10%.

The 2× ETF daily return is calculated as follows:

$= -2 \times [(99 - 110)/110]$

$= -2 \times (-10\%)$

$= 20\%.$

The −2× ETF NAV is calculated as follows:

= 80 × (1 + 0.2).

= 96.

This example shows the fund delivering its promised performance, −2× the daily index return, but it also shows how the return may not be what is naively expected over periods longer than a day.

Over the three days, the index return is −1%: (99 − 100)/100. A naive expectation might assume that over the same period, the −2× ETF would return 2% (= −2 × −1%). Over the three days, the fund's actual return was −4%: (96 − 100)/100.

Because of these compounding effects in leveraged ETFs, the funds are generally not intended to be buy-and-hold products for more than a one-month horizon. If investors are planning to hold them long term, they must rebalance the funds periodically to maintain the desired net exposure.

5 ETFS IN PORTFOLIO MANAGEMENT

☐ identify and describe portfolio uses of ETFs

ETFs have become valuable tools for both institutional and retail investors. Available on a wide range of passive, systematic (rules-based) active, and traditional active strategies and segments of the stock, bond, and commodity markets, ETFs are used for both top-down (based on macro views) and bottom-up (focused on security selection) investment approaches. In addition to their use in implementing long-term strategic exposure to asset classes and risk factors, ETFs are used for tactical tilts, portfolio rebalancing, and risk management.

ETF Strategies

Most institutional asset managers and hedge fund managers, Registered Investment Advisers (RIAs), and financial advisers use ETFs for a wide range of strategies. These strategies serve many different investment objectives—some strategic, some tactical, and some dynamic, where the timing of changes is based on market conditions. Other ETF applications help in managing portfolios more efficiently and are used primarily for operational purposes. As we discuss the diverse set of strategies that can be found in an ETF structure, it is apparent that they are not easily classified as either active or passive. Except for core asset class and portfolio efficiency investment applications that use ETFs based on market-capitalization weighted benchmarks, almost all ETF-related strategies have some component of active investing, either within the ETF strategy or in the way the ETF is used.

Not all strategies are suitable in an ETF structure. The disclosure of holdings may be undesirable for an active manager who invests in less liquid securities or pursues either a concentrated investment strategy or one that relies on an approach that cannot be easily described (such as a "black box" methodology) or disclosed without compromising the strategy. The liquidity of the underlying investments must also be high enough to accommodate daily creations and redemptions. Such factors as tax efficiency, low fees, and available product make ETFs competitive alternatives to traditional mutual funds and active managers. The primary applications in which ETFs are used include the following:

Portfolio efficiency: The use of ETFs to better manage a portfolio for efficiency or operational purposes. Applications include cash or liquidity management, rebalancing, portfolio completion, and active manager transition management.

Asset class exposure management: The use of ETFs to achieve or maintain core exposure to key asset classes, market segments, or investment themes on a strategic, tactical, or dynamic basis.

Active and factor investing: The use of ETFs to target specific active or factor exposures on the basis of an investment view or risk management need.

Efficient Portfolio Management

ETFs are useful tools for managing portfolio activity necessitated by cash flows and changes in external managers. In addition, ETFs can be used to easily accommodate portfolio rebalancing needs and unwanted gaps in portfolio exposure.

Portfolio liquidity management.

One of the primary institutional applications of ETFs is cash flow management. ETFs can be used to invest excess cash balances quickly (known as cash equitization), enabling investors to remain fully invested in target benchmark exposure, thereby minimizing potential cash drag. Cash drag refers to a fund's mis-tracking relative to its index that results from holding uninvested cash. Managers may also use ETFs to transact small cash flows originating from dividends, income, or shareholder activity. Some portfolio managers hold small portions of their funds in ETFs in anticipation of future cash outflows. Transacting the ETF may incur lower trading costs and be easier operationally than liquidating underlying securities or requesting funds from an external manager.

Portfolio rebalancing.

Many investors rebalance portfolios on the basis of a specified time interval, usually at least quarterly, and some may adjust whenever the market value of a portfolio segment, or allocation, deviates from its target weight by a threshold, such as 2%. For tighter rebalancing thresholds and more frequent rebalancing time intervals, using liquid ETFs with tight bid–ask spreads allows the portfolio manager to execute the rebalance in a single ETF trade and ensures the portfolio remains fully invested according to its target weights. For investors who have the ability to sell short, reducing exposure associated with a rebalance can be done quickly using an ETF, and as the underlying securities are sold off, the short position can be covered.

Portfolio completion strategies.

ETFs can also be used for completion strategies to fill a temporary gap in exposure to an asset class, sector, or investment theme or factor. Gaps may arise with changes in external managers or when an existing manager takes an active view that moves the portfolio out of a market segment to which the investor wishes to have continued exposure. The investor may want to retain the manager but use a tactical ETF strategy to maintain exposure to the desired market segment. If external managers are collectively underweighting or overweighting an industry or segment, such as technology, international small-cap stocks, or high-yield bonds, ETFs can be used to adjust exposure up or down to the desired level without making changes to underlying external manager allocations.

Transition management.

Transition management refers to the process of hiring and firing managers—or making changes to allocations with existing managers—while trying to keep target allocations in place. Because ETFs exist on most domestic, international, and global equity benchmarks, a newly appointed transition manager can invest in an ETF to maintain market exposure as she undergoes the process of selling the unwanted positions of the manager she is replacing (the terminated manager). The new transition manager can then take her time to invest in positions for her strategy and gradually reduce the ETF holding.

Asset owners can use ETFs to maintain desired market or asset class exposure in the absence of having an external manager in place. For example, if a fixed-income manager benchmarked to the Bloomberg Barclays US Aggregate Bond Index is terminated, the asset owner may wish to invest in the iShares Core US Aggregate Bond ETF (AGG) to maintain benchmark exposure until a replacement manager can be hired. In some cases, asset owners will "fund" new managers with ETF positions. The new manager will then sell off his ETF positions in the benchmark index as he invests in the underlying securities that meet his desired investment objectives and valuation criteria.

For very large asset owners, there are three potential drawbacks to using ETFs for portfolio management: (1) Given the asset owner size, they may be able to negotiate lower fees for a dedicated separately managed account (SMA) or find lower-cost commingled trust accounts that offer lower fees for large investors, (2) an SMA can be customized to the investment goals and needs of the investor, and (3) many regulators require large ETF holdings (as a percentage of ETF assets) to be disclosed to the public. This can detract from the flexibility in managing the ETF position and increase the cost of shifting investment holdings.

Exhibit 16 provides a summary of ETF portfolio efficiency applications, covering their roles in the portfolio, and examples by benchmark type. Applications include (1) transacting cash flows for benchmark exposure, (2) rebalancing to target asset class or risk factor weights, (3) filling exposure gaps in portfolio holdings of other strategies and funds, and (4) temporarily holding during transitions of strategies or managers.

Exhibit 16: ETF Portfolio Applications—Portfolio Efficiency

Portfolio Application	Role in Portfolio	Examples of ETFs by Benchmark Type
Cash Equitization/ Liquidity Management	Minimize cash drag by staying fully invested to benchmark exposure, transact small cash flows	Liquid ETFs benchmarked to asset category
Portfolio Rebalancing	Maintain exposure to target weights (asset classes, sub-asset classes)	Domestic equity, international equity, domestic fixed income
Portfolio Completion	Fill gaps in strategic exposure (countries, sectors, industries, themes, factors)	International small cap, Canada, bank loans, real assets, health care, technology, quality, ESG
Manager Transition Activity	Maintain interim benchmark exposure during manager transitions	ETFs benchmarked to new manager's target benchmark

Asset Class Exposure Management

Investors have used index exposure in core asset classes for decades, but one of the fastest-growing areas of ETF usage, especially by institutional investors, is fixed income. Since the financial crisis of 2008, the reduced capital available for banks (to participate in dealer bond markets) has contributed to greater use of fixed-income ETFs for core exposure. Except for the largest institutional investors, trading portfolios of bonds is much more difficult and expensive than similar portfolio trades in stocks. Fixed-income ETFs, especially those benchmarked to indexes containing corporates and high-yield securities, provide bond investors with a more efficient (lower cost, more continuous pricing, agency market) and liquid means of obtaining core fixed-income exposure.

Core exposure to an asset class or sub-asset class.

The primary strategic use of ETFs is to gain core index exposure to various asset classes and sub-asset classes. ETFs make doing so easy—across global equities, bonds, commodities, and currencies—and investors regularly use ETFs for broad portfolio diversification. Investors also use ETFs for more targeted strategic exposure to such segments as high-yield debt, bank loans, and commodities (including crude oil, gold and other metals, and agricultural products).

A financial adviser can use ETFs to build a diversified portfolio on the basis of ETF recommendations from his firm's wealth management research team. Benchmarked to broad asset classes, portfolio choices for equity ETF exposure might include domestic large- and small-cap equities, sectors, such risk factors as dividend growth or momentum, industries, and international regions or countries with or without currency exposure. Choices for fixed-income ETF exposure might include government and corporate debt of various maturities, emerging market debt, bank loans, and possibly floating interest rate strategies. Commodity ETF exposure could include gold and other metals, broad commodity indexes, agriculture products, and oil. Similarly, brokerage firms and robo-advisers may offer more-automated solutions that select an ETF allocation based on the investor's risk and return profile. These firms offer a range of ETF investment choices from a preapproved product list to fit different asset class and risk factor categories.

Tactical strategies.

ETFs can also be used to implement market views and adjust portfolio risk on a more short-term, tactical basis. Some financial advisers and institutional investors allocate a portion of their portfolios for opportunistic trading based on their firm's (or strategist's) research or short-term outlook. Others make tactical adjustments in a range around target weights for asset classes or categories within an asset class. ETFs based on risk factors, country exposure, credit or duration exposure, currencies, or even volatility, crude oil, or metals can be used to express tactical views. To profit from an expected price decline, investors can sell ETFs short in a margin account.

Thematic ETFs are also used to implement investment views. Thematic ETFs hold stocks passively but allow investors to take an active view on a market segment they believe will deliver strong returns. These ETFs typically cover a narrow or niche area of the market not well represented by an industry. Examples include focused areas of technology, such as cybersecurity and robotics. Other themes accessed via ETFs are global infrastructure, regional banks, semiconductors, and gold mining. Generally, thematic ETFs are tactical tools that serve as substitutes for buying individual stocks or an industry ETF that is too broad to adequately represent the investor's investment view. Holdings may overlap with those of other ETFs or other portfolio positions but play a role when the investor wants to overweight this segment in the portfolio. Thematic ETFs should be evaluated similarly to stocks because they tend to have comparable levels of volatility and represent specialized active views.

ETFs that have the highest trading volumes in their asset class category are generally preferred for tactical trading applications, and the liquidity in many of the largest ETFs offered in each region makes them well suited for this purpose. Trading costs and liquidity, rather than management fees, are the important criteria in selecting an ETF for tactical adjustments. To identify the most commonly used ETFs for tactical strategies, one can look at the ratio of average dollar volume to average assets for the ETF.

Exhibit 17 provides a summary of ETF asset class exposure applications, covering their roles in the portfolio, categories of use, and examples by benchmark type. These applications relate to using ETFs for strategic, tactical, and dynamic asset class exposure.

Exhibit 17: ETF Portfolio Applications—Asset Class Exposure Management

Portfolio Application	Category	Role in Portfolio	Examples of ETFs by Benchmark Type
Core asset class or market	Strategic or tactical	Core long-term, strategic weighting Tactical tilt to enhance returns or modify risk Ease of access vs. buying underlying securities	Domestic equity, international equity, fixed income, commodities
Equity style, country, or sector; fixed income or commodity segment	Strategic or tactical	Tactical tilt to enhance returns or modify risk depending on short-term views Hedge index exposure of active stocks or bond strategy Ease of access vs. buying underlying securities	Value, growth, Japanese, Chinese, UK, Canadian, or Mexican equities; corporate or high-yield debt; gold; oil; agriculture
Equity sector, industry, investment theme	Dynamic or tactical	Tactical or dynamic active tilt to enhance returns or modify risk Efficient implementation of a thematic/industry vs. single-stock view Capture performance on an emerging theme or innovation not reflected in industry categories	Technology, financials, oil and gas, biotech, infrastructure, robotics, gold mining, buybacks, internet innovation, cybersecurity

Active and Factor Investing

In the mid-2000s, quantitative or rules-based strategies became available in ETFs. These strategies had "active" weights different from market capitalization and were able to disclose holdings because the stock selection and weighting was not chosen by a discretionary portfolio manager but, rather, by a set of quantitative rules, disclosed in the index methodology.

The first smart beta ETFs were indexes weighted by company fundamentals, such as dividends, or quantitatively screened on stock features. Although adoption was initially slow, institutional investors and RIAs now use smart beta ETF strategies to gain systematic active exposure to persistent common return drivers or factors. Global assets in smart beta equity funds, including both single-factor and multi-factor strategies, now represent approximately 20% of ETF assets.

Active ETFs, where the investment strategy is benchmarked but managed with discretion, have also gained assets, especially in fixed income, but they still represent a relatively small percentage of global ETF assets, at 2%–3%.

Factor (smart beta) ETFs.

Factor ETFs are usually benchmarked to an index created with predefined rules for screening and/or weighting constituent holdings. The strategy index rules are structured around return drivers or factors, such as value, dividend yield, earnings or dividend growth, quality, stock volatility, or momentum. Some of these factors, such as size, value, and momentum, have academic support as equity risk premiums that may be rewarded over the long term. Within each single factor category, a range of offerings from competing ETF providers exists, differentiated by the criteria used to represent the factor and the weights applied to constituent holdings (equal, factor, or cap weighted). Their application is typically in providing longer-term, buy-and-hold exposure to a desired factor based on an investment view. Factor ETFs can be used to add risk factor allocations that might not be present in a benchmark or portfolio—for example, adding an equity index ETF with stocks screened for quality to add desired exposure to a quality factor.

Multi-factor ETFs that combine several factors also exist. They may adjust their weights dynamically as market opportunities and risk change. In a multi-factor ETF, strategy design involves factor selection, factor strategy construction, and a weighting scheme across factors that is managed over time. A multi-factor approach typically has lower return volatility than a single-factor approach over time but may also have less return potential for investors who want to capitalize on factor timing.

The success of active strategy ETFs is related to (1) whether the factor, as represented by a target benchmark factor index, performs well relative to expectations and (2) how effective the selected ETF is at delivering the benchmark factor return. Just as with traditional active investing, the success of active investing with ETFs depends on the skill of the ETF portfolio manager as well as the end investor's decision to undertake the investment strategy.

Risk management.

Some smart beta ETFs are constructed to deliver lower or higher risk than that of their asset class benchmark. For example, low-volatility factor ETFs select stocks on the basis of their relative return volatility and seek to represent a portfolio that offers a lower or target volatility return profile. These low-volatility rules-based factor ETFs have gained assets within each segment of the global equity market (domestic, developed international, and emerging markets) as investors have moved to lower volatility in portfolios. Other ETFs based on the beta characteristics of the constituent stocks can be used to adjust the portfolio's beta profile to desired levels.

ETFs are also used to manage other portfolio risks, such as currency and duration risk. ETFs that provide international exposure with a hedge on all or part of the associated currency risk are available. With respect to interest rate risk management, several smart beta fixed-income ETFs hold long positions in corporate or high-yield bonds and hedge out the duration risk of these bonds with futures or short positions in government bonds. These ETFs enable investors to add a position to their portfolio that seeks returns from taking credit risk with minimal sensitivity to movements in interest rates. Active investors with a negative macro view can use inverse asset class or factor ETF exposure to temporarily reduce benchmark holding risk. Doing so allows them to implement a macro view on a short-term basis and minimize turnover in underlying portfolio holdings.

Alternatively weighted ETFs.

ETFs that weight their constituents by means other than market capitalization, such as equal weighting or weightings based on fundamentals, can also be used to implement investment views—for example, ETFs that weight constituent stocks on the basis of their dividend yields. These ETFs select or overweight stocks with higher dividend yields, subject to other fundamental criteria or constraints, and are used by investors seeking income-generating strategies.

Discretionary active ETFs.

The largest active ETFs are in fixed income, where passive management is much less prominent than in equities. The PIMCO Active Bond ETF (BOND) launched in 2012 with an investment objective similar to that of the world's largest mutual fund at that time, the PIMCO Total Return Fund. Shorter-maturity, actively managed ETFs are also available in fixed income. Other active ETFs include exposure to senior bank loans, floating rate debt, and mortgage securities. Active equity ETFs have also been launched in areas of the technology industry.

"Liquid alternative" ETFs are based on strategy indexes that attempt to deliver absolute return performance and/or risk diversification of stock and bond holdings. Some of the first liquid alternative ETFs used rules-based strategies to replicate broad hedge fund indexes. Other strategy indexes offer transparent, rules-based, "hedge fund–like" strategies in specific types of alternatives. Such strategies include long–short, managed futures, private equity, and merger arbitrage.

Dynamic asset allocation and multi-asset strategies.

ETF availability across a wide range of equity and bond risk exposures has fostered greater use of dynamic, top-down investment strategies based on return and risk forecasts. Asset managers, hedge funds, and asset owners have increasingly used ETFs for discretionary asset allocation or global macro strategies. Dynamic asset allocation ETF strategies are also available in commodities. Although some strategies allocate holdings on the basis of their relative risk contribution and others are return focused, all involve adjustments back to target weights, as defined by a dynamic investment process. Some pension and sovereign wealth funds implement these strategies in house, whereas other investors hire asset managers that offer multi-asset strategies. Implementation is done using ETFs, along with futures and swaps where available and when they are more efficient to trade.

Proper use of an active or factor strategy ETF requires investors to research and assess the index construction methodology and performance history and to ensure consistency with their investment view.

Exhibit 18 provides a summary of active and factor ETF portfolio applications, covering their roles in the portfolio, categories of use, and examples by benchmark type. These applications relate to ETFs as alternatives to other fund products, such as active mutual funds. In these cases, ETF evaluation is based on features of the investment approach, holdings, cost, risk, and return potential, as well as the impact to the portfolio's overall risk and return.

Exhibit 18: ETF Portfolio Applications—Active and Factor Investing

Portfolio Application	Category	Role in Portfolio	Examples of ETFs by Benchmark Type
Factor exposure	Strategic, dynamic, or tactical	Capture risk premium for one or more factors driving returns or risk Overweight or underweight depending on factor return or risk outlook Seek to capture alpha from rules-based screening and rebalancing (systematic active)	Quality, dividend growth, value, momentum, low volatility, liquidity screen, multi-factor
Risk management	Dynamic or tactical	Adjust equity beta, duration, credit, or currency risk	Currency-hedged, low-volatility, or downside-risk-managed ETFs
Leveraged and inverse exposure	Tactical	Access leveraged or short exposure for short-term tilts or risk management Limit losses on shorting to invested funds	ETFs representing asset classes, countries, or industries with leveraged or inverse daily return targets
Alternative weighting	Strategic, dynamic, or tactical	Seek outperformance from weighting based on one or more fundamental factors Balance or manage risk of security holdings	ETFs weighted by fundamentals, dividends, or risk; equal-weighted ETFs

Portfolio Application	Category	Role in Portfolio	Examples of ETFs by Benchmark Type
Active strategies within an asset class	Strategic	Access discretionary active management in an ETF structure	ETFs from reputable fixed income or equity managers with active approach or theme
Dynamic asset allocation and multi-asset strategies	Dynamic or tactical	Seek returns from active allocation across asset classes or factors based on return or risk outlook Invest in a multi-asset-class strategy in single product	ETFs that allocate across asset categories or investment themes based on quantitative or fundamental factors

SUMMARY

We have examined important considerations for ETF investors, including how ETFs work and trade, tax efficient attributes, and key portfolio uses. The following is a summary of key points:

- ETFs rely on a creation/redemption mechanism that allows for the continuous creation and redemption of ETF shares.
- The only investors who can create or redeem new ETF shares are a special group of institutional investors called authorized participants.
- ETFs trade on both the primary market (directly between APs and issuers) and on the secondary markets (exchange-based or OTC trades, such as listed equity).
- End investors trade ETFs on the secondary markets, like stocks.
- Holding period performance deviations (tracking differences) are more useful than the standard deviation of daily return differences (tracking error).
- ETF tracking differences from the index occur for the following reasons:
  - fees and expenses,
  - representative sampling/optimization,
  - use of depositary receipts and other ETFs,
  - index changes,
  - fund accounting practices,
  - regulatory and tax requirements, and
  - asset manager operations.
- ETFs are generally taxed in the same manner as the securities they hold, with some nuances:
  - ETFs are more tax fair than traditional mutual funds, because portfolio trading is generally not required when money enters or exits an ETF.
  - Owing to the creation/redemption process, ETFs can be more tax efficient than mutual funds.
  - ETF issuers can redeem out low-cost-basis securities to minimize future taxable gains.
  - Local markets have unique ETF taxation issues that should be considered.

- ETF bid–ask spreads vary by trade size and are usually published for smaller trade sizes. They are tightest for ETFs that are very liquid and have continuous two-way order flow. For less liquid ETFs, the following factors can determine the quoted bid–ask spread of an ETF trade:
    - Creation/redemption costs, brokerage and exchange fees
    - Bid–ask spread of underlying securities held by the ETF
    - Risk of hedging or carry positions by liquidity provider
    - Market makers' target profit spread
- ETF bid–ask spreads on fixed income relative to equity tend to be wider because the underlying bonds trade in dealer markets and hedging is more difficult. Spreads on ETFs holding international stocks are tightest when the underlying security markets are open for trading.
- ETF premiums and discounts refer to the difference between the exchange price of the ETF and the fund's calculated NAV, based on the prices of the underlying securities and weighted by the portfolio positions at the start of each trading day. Premiums and discounts can occur because NAVs are based on the last traded prices, which may be observed at a time lag to the ETF price, or because the ETF is more liquid and more reflective of current information and supply and demand than the underlying securities in rapidly changing markets.
- Costs of ETF ownership may be positive or negative and include both explicit and implicit costs. The main components of ETF cost are
    - the fund management fee;
    - tracking error;
    - portfolio turnover;
    - trading costs, such as commissions, bid–ask spreads, and premiums/discounts;
    - taxable gains/losses; and
    - security lending.
- Trading costs are incurred when the position is entered and exited. These one-time costs decrease as a portion of total holding costs over longer holding periods and are a more significant consideration for shorter-term tactical ETF traders.
- Other costs, such as management fees and portfolio turnover, increase as a proportion of overall cost as the investor holding period lengthens. These costs are a more significant consideration for longer-term buy-and-hold investors.
- ETFs are different from exchange-traded notes, although both use the creation/redemption process.
    - Exchange-traded notes carry unique counterparty risks of default.
    - Swap-based ETFs may carry counterparty risk.
    - ETFs, like mutual funds, may lend their securities, creating risk of counterparty default.
    - ETF closures can create unexpected tax liabilities.

- ETFs are used for core asset class exposure, multi-asset, dynamic, and tactical strategies based on investment views or changing market conditions; for factor or smart beta strategies with a goal to improve return or modify portfolio risk; and for portfolio efficiency applications, such as rebalancing, liquidity management, completion strategies, and transitions.
- ETFs are useful for investing cash inflows, as well as for raising proceeds to provide for client withdrawals. ETFs are used for rebalancing to target asset class weights and for "completion strategies" to fill a temporary gap in an asset class category, sector, or investment theme or when external managers are underweight. When positions are in transition from one external manager to another, ETFs are often used as the temporary holding and may be used to fund the new manager.
- All types of investors use ETFs to establish low-cost core exposure to asset classes, equity style benchmarks, fixed-income categories, and commodities.
- For more tactical investing, thematic ETFs are used in active portfolio management and represent narrow or niche areas of the equity market not well represented by industry or sector ETFs.
- Systematic, active strategies that use rules-based benchmarks for exposure to such factors as size, value, momentum, quality, or dividend tilts or combinations of these factors are frequently implemented with ETFs.
- Multi-asset and global asset allocation or macro strategies that manage positions dynamically as market conditions change are also areas where ETFs are frequently used.
- Proper utilization requires investors to carefully research and assess the ETF's index construction methodology, costs, risks, and performance history.

PRACTICE PROBLEMS

1. Which of the following statements regarding exchange-traded funds (ETFs) is correct? ETFs:
   - **A.** disclose their holdings on a quarterly basis.
   - **B.** trade in both primary and secondary markets.
   - **C.** offer a creation/redemption mechanism that allows any investor to create or redeem shares.

2. The list of securities that a particular ETF wants to own, which is disclosed daily by all ETFs, is referred to as the:
   - **A.** creation unit.
   - **B.** creation basket.
   - **C.** redemption basket.

3. When an authorized participant transacts to create or redeem ETF shares, the related costs are ultimately borne:
   - **A.** solely by the ETF sponsor.
   - **B.** by transacting shareholders.
   - **C.** proportionally by all existing ETF shareholders.

4. Assuming arbitrage costs are minimal, which of the following is *most likely* to occur when the share price of an ETF is trading at a premium to its intraday NAV?
   - **A.** New ETF shares will be created by the ETF sponsor.
   - **B.** Redemption baskets will be received by APs from the ETF sponsor.
   - **C.** Retail investors will exchange baskets of securities that the ETF tracks for creation units.

The following information relates to questions 5-10

Howie Rutledge is a senior portfolio strategist for an endowment fund. Rutledge meets with recently hired junior analyst Larry Stosur to review the fund's holdings.

Rutledge asks Stosur about the mechanics of exchange-traded funds (ETFs). Stosur responds by making the following statements:

Statement 1	Unlike mutual fund shares that can be shorted, ETF shares cannot be shorted.
Statement 2	In the ETF creation/redemption process, the authorized participants (APs) absorb the costs of transacting securities for the ETF's portfolio.

Statement 3 If ETF shares are trading at a discount to NAV and arbitrage costs are sufficiently low, APs will buy the securities in the creation basket and exchange them for ETF shares from the ETF sponsor.

Rutledge notes that one holding, ETF 1, is trading at a premium to its intraday NAV. He reviews the ETF's pricing and notes that the premium to the intraday NAV is greater than the expected arbitrage costs.

Stosur is evaluating three ETFs for potential investment. He notes that the ETFs have different portfolio characteristics that are likely to affect each ETF's tracking error. A summary of the characteristics for the ETFs is presented in Exhibit 1.

Exhibit 1: ETF Characteristics Affecting Tracking Error

	ETF 2	ETF 3	ETF 4
Portfolio Construction Approach	Full Replication	Representative Sampling	Full Replication
Type of Foreign Holdings	Local shares	ADRs*	ADRs*
Engagement in Securities Lending	Yes	Yes	No

**ADRs are American Depositary Receipts.*

Rutledge and Stosur discuss the factors that influence ETF bid–ask spreads. Stosur tells Rutledge that quoted bid–ask spreads for a particular transaction size are (1) negatively related to the amount of the ongoing order flow in the ETF, (2) positively related to the costs and risks for the ETF liquidity provider, and (3) positively related to the amount of competition among market makers for the ETF.

As ETF shares may trade at prices that are different from the NAV, Rutledge examines selected data in Exhibit 2 for three ETFs that might have this problem.

Exhibit 2: Selected Data on ETFs

	ETF 5	ETF 6	ETF 7
Percentage of Foreign Holdings	10%	50%	90%
Trading Frequency	High	Low	Low

Rutledge considers a new ETF investment for the fund. He plans to own the ETF for nine months. The ETF has the following trading costs and management fees:

- Annual management fee of 0.32%
- Round-trip trading commissions of 0.20%
- Bid–offer spread of 0.10% on purchase and sale

Rutledge asks Stosur to compute the expected total holding period cost for investing in the ETF.

5. Which of Stosur's statements regarding ETF mechanics is correct?

   A. Statement 1

   B. Statement 2

   C. Statement 3

6. Given the current pricing of ETF 1, the *most likely* transaction to occur is that:
   - **A.** new ETF shares will be created by the APs.
   - **B.** redemption baskets will be received by APs from the ETF sponsor.
   - **C.** retail investors will exchange baskets of securities that the ETF tracks for creation units.

7. Which ETF in Exhibit 1 is *most likely* to have the lowest tracking error?
   - **A.** ETF 2
   - **B.** ETF 3
   - **C.** ETF 4

8. Stosur's statement about quoted bid–ask spreads is *incorrect* with respect to the:
   - **A.** amount of the ongoing order flow in the ETF.
   - **B.** costs and risks for the ETF liquidity providers.
   - **C.** amount of competition among market makers for the ETF.

9. Which ETF in Exhibit 2 is *most likely* to trade at the largest premium or discount relative to NAV?
   - **A.** ETF 5
   - **B.** ETF 6
   - **C.** ETF 7

10. Excluding the compounding effect, the expected total holding period cost for investing in the ETF over a nine-month holding period is *closest* to:
    - **A.** 0.54%.
    - **B.** 0.62%.
    - **C.** 0.64%.

---

11. An ETF's reported tracking error is typically measured as the:
    - **A.** standard deviation of the difference in daily returns between an ETF and its benchmark.
    - **B.** difference in annual return between an ETF and its benchmark over the past 12 months.
    - **C.** annualized standard deviation of the difference in daily returns between an ETF and its benchmark.

12. To best assess an ETF's performance, which reflects the impact of portfolio rebalancing expenses and other fees, an investor should:
    - **A.** review daily return differences between the ETF and its benchmark.
    - **B.** perform a rolling return assessment between the ETF and its benchmark.

C. compare the ETF's annual expense ratio with that of other ETFs in its asset class category.

13. An ETF's tracking error, as traditionally reported, indicates to investors:

A. whether the ETF is underperforming or outperforming its underlying index.

B. the magnitude by which an ETF's returns deviate from its benchmark over time.

C. the distribution of differences in daily returns between the ETF and its benchmark.

14. For a typical ETF, which of the following sources of tracking error is *most likely* to be the smallest contributor to tracking error?

A. Representative sampling

B. Fees and expenses incurred by the ETF

C. Changes to the underlying index securities

15. Which of the following statements relating to capital gains in ETFs and mutual funds is correct?

A. ETFs tend to distribute less in capital gains than mutual funds do.

B. Mutual funds may elect not to distribute all realized capital gains in a given year.

C. The selling of ETF shares by some investors may create capital gains that affect the remaining ETF investors in terms of taxes.

16. Which of the following statements regarding distributions made by ETFs is correct?

A. Return-of-capital (ROC) distributions are generally not taxable.

B. ETFs generally reinvest any dividends received back into the ETF's holdings.

C. A dividend distribution is a distribution paid to investors in excess of an ETF's earnings.

17. Investors buying ETFs:

A. incur management fees that decrease with the length of the holding period.

B. are assured of paying a price equal to the NAV if they purchase shares at the market close.

C. incur trading costs in the form of commissions and bid–ask spreads at the time of purchase.

18. Consider an ETF with the following trading costs and management fees:

- Annual management fee of 0.40%
- Round-trip trading commissions of 0.55%

- Bid–offer spread of 0.20% on purchase and sale

Excluding compound effects, the expected total holding-period cost for investing in the ETF over a nine-month holding period is *closest* to:

A. 1.05%.

B. 1.15%.

C. 1.25%.

19. The bid–ask spread for very liquid, high-volume ETFs will be *least* influenced by the:

A. market maker's desired profit spread.

B. creation/redemption fees and other direct costs.

C. likelihood of receiving an offsetting ETF order in a short time frame.

20. Such factors as regulations, competition, and corporate actions relate to:

A. fund-closure risk.

B. counterparty risk.

C. expectation-related risk.

21. John Smith has invested in an inverse ETF. Smith is a novice investor who is not familiar with inverse ETFs, and therefore, he is unsure how the ETF will perform because of a lack of understanding of the ETF's risk and return characteristics. This risk is *best* described as:

A. counterparty risk.

B. holdings-based risk.

C. expectation-related risk.

22. Factor (smart beta) strategy ETFs are *least likely* to be used by investors:

A. to modify portfolio risk.

B. for tactical trading purposes.

C. to seek outperformance versus a benchmark.

23. Which of the following statements regarding applications of ETFs in portfolio management is correct?

A. Equity ETFs tend to be more active than fixed-income ETFs.

B. The range of risk exposures available in the futures market is more diverse than that available in the ETF space.

C. ETFs that have the highest trading volumes in their asset class category are generally preferred for tactical trading applications.

SOLUTIONS

1. B is correct. ETFs trade in both primary and secondary markets. The primary market for ETF trading is that which exists on an over-the-counter basis between authorized participants (APs), a special group of institutional investors, and the ETF issuer or sponsor. This process is referred to as creation/redemption, and it is only through these primary market transactions that shares of the ETF can be created or destroyed. ETFs also trade in the secondary market on exchanges. Secondary market trading happens between any pair of market participants—individual or institutional investors, market makers, and so on.

2. B is correct. Each day, ETF managers publicly disclose a list of securities that they want to own, which is referred to as the creation basket. This basket also serves as the portfolio for determining the intrinsic net asset value (NAV) of the ETF on the basis of prices during the trading day.

3. B is correct. The AP generally absorbs all the costs associated with buying or selling the securities in the baskets or the ETF shares and pays an additional fee to the ETF provider to cover processing fees associated with creation/redemption activities. APs pass these costs to investors in the ETF's bid–ask spread, which is incurred by investors entering (ETF share buyers) and exiting (ETF share sellers) the fund.

4. A is correct. When the share price of an ETF is trading at a premium to its intraday NAV and assuming arbitrage costs are minimal, APs will step in and take advantage of the arbitrage. Specifically, APs will step in and buy the basket of securities that the ETF tracks (the creation basket) and exchange it with the ETF provider for new ETF shares (a creation unit). These new shares received by APs can then be sold on the open market to realize arbitrage profits.

5. B is correct. Statement 2 is correct. A significant advantage of the ETF creation/redemption process is that the AP absorbs all costs of transacting the securities for the fund's portfolio. APs pass these costs to investors in the ETF's bid–ask spread, incurred by ETF buyers and sellers. Thus, non-transacting shareholders of an ETF are shielded from the negative impact of transaction costs caused by other investors entering and exiting the fund. In contrast, when investors enter or exit a traditional mutual fund, the mutual fund manager incurs costs to buy or sell investments arising from this activity, which affects all fund shareholders. This makes the ETF structure inherently fairer: Frequent ETF traders bear the cost of their activity, while buy-and-hold ETF shareholders are shielded from those costs. Investors cannot short mutual fund shares, but they can short ETF shares. Also, if ETF shares are trading at a discount to NAV and arbitrage costs are sufficiently low, APs will buy ETF shares and exchange them for the securities in the redemption basket. Statement 3 describes the scenario that would occur if the ETF shares are trading at a premium to NAV.

   A is incorrect because Statement 1 is incorrect. Investors cannot short mutual fund shares, but they can short ETF shares.

   C is incorrect because Statement 3 is incorrect. If ETF shares are trading at a discount to NAV and arbitrage costs are sufficiently low, APs will buy ETF shares and exchange them for the securities in the redemption basket. Statement 3 describes the scenario that would occur if ETF shares are trading at a premium to NAV.

6. A is correct. When the share price of an ETF is trading at a premium to its intr-

aday NAV and arbitrage costs are minimal, APs will step in and take advantage of the arbitrage. Specifically, APs will buy the basket of securities that the ETF tracks (the creation basket) and exchange it with the ETF sponsor for new ETF shares (a creation unit). These new ETF shares received by APs can then be sold on the open market to realize arbitrage profits.

B is incorrect because in the case of an ETF trading at a premium to NAV, the APs will not receive redemption baskets of securities. Instead, the APs will deliver creation baskets to the ETF sponsor and receive new ETF shares.

C is incorrect because only APs can deliver creation baskets or receive redemption baskets from the ETF sponsors. Retail investors can buy and sell ETF shares on the open market.

7. A is correct. Compared with a full replication approach, ETF portfolios managed using a representative sampling/optimization approach are likely to have greater tracking error. Also, differences in trading hours for depositary receipts and local constituent shares create discrepancies between the portfolio and index values. These discrepancies can lead to greater tracking error for portfolios holding ADRs in lieu of the underlying local shares. In exchange, the ETF receives a fee and earns interest on the collateral posted by the borrower (generally, overnight fixed-income securities), which creates income for the portfolio. As long as this securities-lending income is less than the fund expenses, it will lower the tracking error, but if it becomes larger, then it will become a source of tracking error. ETF 2 uses a full replication approach, holds only local foreign shares, and engages in securities lending. Therefore, ETF 2 will likely have the lowest tracking error out of the ETFs in Exhibit 1. ETF 3 will likely have greater tracking error than ETF 2 because it is managed using a representative sampling approach and is invested in depositary receipts in lieu of local shares. ETF 4 will likely have greater tracking error than ETF 2 because it is invested in depositary receipts in lieu of local shares and does not engage in securities lending.

8. C is correct. Several factors determine the width of an ETF's quoted bid–ask spread. First, the amount of ongoing order flow in the ETF is negatively related to the bid–ask spread (more flow means lower spreads). Second, the actual costs and risks for the liquidity provider are positively related to spreads (more costs and risks mean higher spreads); the spread is compensation to the liquidity provider for incurring these costs and risks. Finally, the amount of competition among market makers for that ETF is negatively related to the bid–ask spread (more competition means lower spreads).

   A is incorrect because Stosur is correct in stating that the quoted bid–ask spread for a particular transaction size is negatively related to the amount of the ongoing order flow in the ETF (more flow means lower spreads).

   B is incorrect because Stosur is correct in stating that the quoted bid–ask spread for a particular transaction size is positively related to the costs and risks for the ETF liquidity provider (more costs and risks mean higher spreads). The bid–ask spread represents the market maker's price for taking the other side of the ETF transaction, which includes the costs and risks to carry the position on its books and/or to hedge the position using underlying securities or closely related ETFs or derivatives.

9. C is correct. ETFs that trade infrequently may have large premiums or discounts to NAV, because the ETF may not have traded in the hours leading up to the market close and NAV may have significantly risen or fallen during that time because of market movement. Furthermore, NAV is often a poor fair value indicator for ETFs holding foreign securities because of differences in exchange closing times between the underlying (e.g., foreign stocks, bonds, or commodities) and the exchange where the ETF trades. Therefore, ETF 7 is most likely to have the largest

discount or premium because it has a low trading frequency and has the highest percentage of foreign holdings among the three ETFs.

A is incorrect because ETF 5 has the lowest percentage of foreign holdings among the three ETFs and is the one ETF with a high trading frequency. Therefore, relative to ETF 7, with its low trading frequency and high foreign holdings, ETF 5 is likely to trade at smaller premiums or discounts.

B is incorrect because ETF 6 has a lower percentage of foreign holdings than ETF 7. Even though both ETF 6 and ETF 7 have the same low trading frequency, the lower percentage of foreign holdings for ETF 6 is likely to result in it trading at smaller premiums or discounts.

10. A is correct. The expected total holding period cost for investing in the ETF over the nine-month holding period is calculated as follows:

Total expected holding period cost

= Annual management fee + Round-trip trading commissions + Bid–offer spread on purchase/sale.

Total expected holding period cost = (9/12) × (0.32%) + 0.20% + 0.10% = 0.54%.

11. C is correct. An ETF's tracking error is typically reported as the annualized standard deviation of the daily differential returns of the ETF and its benchmark.

12. B is correct. A rolling return assessment, referred to in the ETF industry as the "tracking difference," provides a more informative picture of the investment outcome for an investor in an ETF. Such an analysis allows investors to see the cumulative effect of portfolio management and expenses over an extended period. It also allows for comparison with other annual metrics such as a fund's expense ratio. Tracking error, as a statistic, reveals only ETF tracking variability; it does not reveal to investors whether the fund is over- or underperforming its index or whether that tracking error is concentrated over a few days or is more consistently experienced. An ETF's expense ratio does not fully reflect the investor experience. That is, the expense ratio does not reflect the cost of portfolio rebalancing or other fees, making it an inferior assessment measure relative to a rolling return assessment.

13. B is correct. An ETF's tracking error is typically reported as the annualized standard deviation of the daily differential returns of the ETF and its benchmark. Therefore, an ETF's reported tracking error indicates to investors the magnitude by which an ETF's returns deviate from those of its benchmark over time.

14. C is correct. Although additions and deletions of securities from the underlying benchmark index may occur and result in tracking error, such index changes generally occur infrequently (often quarterly). In addition, ETF portfolio managers may work with APs for index rebalance trades to ensure market-on-close pricing to minimize this source of tracking error. Therefore, the resulting tracking error caused by index changes will not likely be as large as the tracking error caused by representative sampling or by fees and expenses incurred by the ETF.

15. A is correct. ETFs tend to distribute far less in capital gains relative to mutual funds. This is mostly due to the fact that ETFs have historically had significantly lower turnover than mutual funds have had.

16. A is correct. Return-of-capital distributions are amounts paid out in excess of an ETF's earnings and serve to reduce an investor's cost basis by the amount of the distribution. These distributions are generally not taxable.

17. C is correct. ETF trading costs in the form of commissions and bid–ask spreads are paid by investors buying or selling ETF shares on an exchange. These trading costs are influenced by the bid–ask spread of the ETF, the size of the trade relative to the normal trading activity of the ETF, and the ease of hedging the ETF by the market-making community. Even the closing price of the ETF on the exchange includes a premium or discount to the NAV, driven by supply and demand factors on the exchange and the market impact costs of executing an exchange transaction. The purchase and sale trading costs of an ETF are paid regardless of holding period, whereas other costs, such as management fees, increase as the holding period lengthens.

18. A is correct. The expected total holding-period cost for investing in the ETF over a nine-month holding period is calculated as follows:

    Total holding-period cost = Annual management fee + Round-trip trading commissions + Bid–offer spread on purchase/sale.

    Total holding-period cost = (9/12) × (0.40%) + 0.55% + 0.20% = 1.05%.

19. B is correct. ETF bid–ask spreads are generally less than or equal to the combination of the following:

    - ± Creation/redemption fees and other direct costs, such as brokerage and exchange fees
    - \+ Bid–ask spread of the underlying securities held by the ETF
    - \+ Compensation for the risk of hedging or carrying positions by liquidity providers (market makers) for the remainder of the trading day
    - \+ Market maker's desired profit spread
    - − Discount related to the likelihood of receiving an offsetting ETF order in a short time frame

    For very liquid and high-volume ETFs, buyers and sellers are active throughout the trading day. Therefore, because most of these ETF trades are matched extremely quickly and never involve the creation/redemption process, the first three factors listed do not contribute heavily to their bid–ask spreads. So, creation/redemption fees and other direct costs are not likely to have much influence on these ETFs' bid–ask spreads.

20. A is correct. Fund-closure risk is the risk that an ETF may shut down. The reasons that lead to an ETF closing down often have to do with changes in regulations, increased competition, and corporate activity (merger and acquisition activity within the ETF industry).

21. C is correct. Expectation-related risk is the risk that some ETF investors may not fully understand how more complex ETFs will perform because of a lack of understanding of sophisticated assets classes and strategies.

22. B is correct. Factor strategy ETFs are usually benchmarked to an index created with predefined rules for screening and/or weighting stock holdings and are considered longer-term, buy-and-hold investment options rather than tactical trading instruments. The strategy index rules are structured around return drivers or factors, such as value, dividend yield, earnings or dividend growth, quality, stock volatility, or momentum. Investors using factor-based investing seek outperformance versus a benchmark or portfolio risk modification.

23. C is correct. ETFs that have the highest trading volumes in their asset class cate-

gory are generally preferred for tactical trading applications.

LEARNING MODULE

4

Using Multifactor Models

by Jerald E. Pinto, PhD, CFA, and Eugene L. Podkaminer, CFA.

Jerald E. Pinto, PhD, CFA, is at CFA Institute (USA). Eugene L. Podkaminer, CFA, is at Franklin Templeton Investments (USA).

LEARNING OUTCOMES

Mastery	*The candidate should be able to:*
☐	describe arbitrage pricing theory (APT), including its underlying assumptions and its relation to multifactor models
☐	define arbitrage opportunity and determine whether an arbitrage opportunity exists
☐	calculate the expected return on an asset given an asset's factor sensitivities and the factor risk premiums
☐	describe and compare macroeconomic factor models, fundamental factor models, and statistical factor models
☐	describe uses of multifactor models and interpret the output of analyses based on multifactor models
☐	describe the potential benefits for investors in considering multiple risk dimensions when modeling asset returns
☐	explain sources of active risk and interpret tracking risk and the information ratio

1 BACKGROUND AND USES

As used in investments, a **factor** is a variable or a characteristic with which individual asset returns are correlated. Models using multiple factors are used by asset owners, asset managers, investment consultants, and risk managers for a variety of portfolio construction, portfolio management, risk management, and general analytical purposes. In comparison to single-factor models (typically based on a market risk factor), multifactor models offer increased explanatory power and flexibility. These comparative strengths of multifactor models allow practitioners to

- build portfolios that replicate or modify in a desired way the characteristics of a particular index;

CFA Institute would like to thank Yin Luo, CPA, PStat, CFA, and Sheng Wang, both of Wolfe Research LLC (USA), for contributing the case "Constructing Multifactor Portfolios" in Section 8.

- establish desired exposures to one or more risk factors, including those that express specific macro expectations (such as views on inflation or economic growth), in portfolios;
- perform granular risk and return attribution on actively managed portfolios;
- understand the comparative risk exposures of equity, fixed-income, and other asset class returns;
- identify active decisions relative to a benchmark and measure the sizing of those decisions; and
- ensure that an investor's aggregate portfolio is meeting active risk and return objectives commensurate with active fees.

Multifactor models have come to dominate investment practice, having demonstrated their value in helping asset managers and asset owners address practical tasks in measuring and controlling risk. We explain and illustrate the various practical uses of multifactor models.

We first describe the modern portfolio theory background of multifactor models. We then describe arbitrage pricing theory and provide a general expression for multifactor models. We subsequently explore the types of multifactor models and certain applications. Lastly, we summarize major points.

Multifactor Models and Modern Portfolio Theory

In 1952, Markowitz introduced a framework for constructing portfolios of securities by quantitatively considering each investment in the context of a portfolio rather than in isolation; that framework is widely known today as modern portfolio theory (MPT). Markowitz simplified modeling asset returns using a multivariate normal distribution, which completely defines the distribution of returns in terms of mean returns, return variances, and return correlations. One of the key insights of MPT is that any value of correlation among asset returns of less than one offers the potential for risk reduction by means of diversification.

In 1964, Sharpe introduced the capital asset pricing model (CAPM), a model for the expected return of assets in equilibrium based on a mean–variance foundation. The CAPM and the literature that developed around it has provided investors with useful and influential concepts—such as alpha, beta, and systematic risk—for thinking about investing. The concept of systematic risk, for example, is critical to understanding multifactor models: An investment may be subject to many different types of risks, but they are generally not equally important so far as investment valuation is concerned. Risk that can be avoided by holding an asset in a portfolio, where the risk might be offset by the various risks of other assets, should not be compensated by higher expected return, according to theory. By contrast, investors would expect compensation for bearing an asset's non-diversifiable risk: **systematic risk**. Theory indicates that only systematic risk should be **priced risk**. In the CAPM, an asset's systematic risk is a positive function of its beta, which measures the sensitivity of an asset's return to the market's return. According to the CAPM, differences in mean return are explained by a single factor: market portfolio return. Greater risk with respect to the market factor, represented by higher beta, is expected to be associated with higher return.

The accumulation of evidence from the equity markets during the decades following the CAPM's development have provided clear indications that the CAPM provides an incomplete description of risk and that models incorporating multiple sources of systematic risk more effectively model asset returns. Bodie, Kane, and Marcus (2017) provide an introduction to the empirical evidence. There are, however, various perspectives in practice on how to model risk in the context of multifactor models. We will examine some of these—focusing on macroeconomic factor models and fundamental factor models—in subsequent sections.

ARBITRAGE PRICING THEORY AND MULTIFACTOR MODELS

2

- ☐ describe arbitrage pricing theory (APT), including its underlying assumptions and its relation to multifactor models
- ☐ define arbitrage opportunity and determine whether an arbitrage opportunity exists
- ☐ calculate the expected return on an asset given an asset's factor sensitivities and the factor risk premiums

In the 1970s, Ross (1976) developed the arbitrage pricing theory (APT) as an alternative to the CAPM. APT introduced a framework that explains the expected return of an asset (or portfolio) in equilibrium as a linear function of the risk of the asset (or portfolio) with respect to a set of factors capturing systematic risk. Unlike the CAPM, the APT does not indicate the identity or even the number of risk factors. Rather, for any multifactor model assumed to generate returns ("return-generating process"), the theory gives the associated expression for the asset's expected return.

Suppose that K factors are assumed to generate returns. Then the simplest expression for a multifactor model for the return of asset i is given by

$$R_i = a_i + b_{i1}I_1 + b_{i2}I_2 + \ldots + b_{iK}I_K + \varepsilon_i, \quad (1)$$

where

R_i = the return to asset i

a_i = an intercept term

I_k = the return to factor k, $k = 1, 2, \ldots, K$

b_{ik} = the sensitivity of the return on asset i to the return to factor k, $k = 1, 2, \ldots, K$

ε_i = an error term with a zero mean that represents the portion of the return to asset i not explained by the factor model

The intercept term a_i is the expected return of asset i given that all the factors take on a value of zero. Equation 1 presents a multifactor return-generating process (a time-series model for returns). In any given period, the model may not account fully for the asset's return, as indicated by the error term. But error is assumed to average to zero. Another common formulation subtracts the risk-free rate from both sides of Equation 1 so that the dependent variable is the return in excess of the risk-free rate and one of the explanatory variables is a factor return in excess of the risk-free rate. (The Carhart model described next is an example.)

Based on Equation 1, the APT provides an expression for the expected return of asset *i* assuming that financial markets are in equilibrium. The APT is similar to the CAPM, but the APT makes less strong assumptions than the CAPM. The APT makes just three key assumptions:

1. A factor model describes asset returns.
2. With many assets to choose from, investors can form well-diversified portfolios that eliminate asset-specific risk.
3. No arbitrage opportunities exist among well-diversified portfolios.

Arbitrage is a risk-free operation that requires no net investment of money but earns an expected positive net profit. (Note that "arbitrage," or the phrase "risk arbitrage," is also sometimes used in practice to describe investment operations in which significant risk is present). An **arbitrage opportunity** is an opportunity to conduct an arbitrage—an opportunity to earn an expected positive net profit without risk and with no net investment of money.

In the first assumption, the number of factors is not specified. The second assumption allows investors to form portfolios with factor risk but without asset-specific risk. The third assumption is the condition of financial market equilibrium.

Empirical evidence indicates that Assumption 2 is reasonable (Fabozzi, 2008). When a portfolio contains many stocks, the asset-specific or non-systematic risk of individual stocks makes almost no contribution to the variance of portfolio returns.

According to the APT, if these three assumptions hold, the following equation holds:

$$E(R_p) = R_F + \lambda_1\beta_{p,1} + \ldots + \lambda_K\beta_{p,K}, \quad (2)$$

where

$E(R_p)$ = the expected return to portfolio p

R_F = the risk-free rate

λ_j = the expected reward for bearing the risk of factor j

$\beta_{p,j}$ = the sensitivity of the portfolio to factor j

K = the number of factors

The APT equation, Equation 2, says that the expected return on any well-diversified portfolio is linearly related to the factor sensitivities of that portfolio. The equation assumes that a risk-free rate exists. If no risk-free asset exists, in place of R_F we write λ_0 to represent the expected return on a risky portfolio with zero sensitivity to all the factors. The number of factors is not specified but must be much lower than the number of assets, a condition fulfilled in practice.

The **factor risk premium** (or **factor price**), λ_j, represents the expected reward for bearing the risk of a portfolio with a sensitivity of 1 to factor *j* and a sensitivity of 0 to all other factors. The exact interpretation of "expected reward" depends on the multifactor model that is the basis for Equation 2. For example, in the Carhart four-factor model, shown later in Equation 3 and Equation 4, the risk premium for the market factor is the expected return of the market in excess of the risk-free rate. Then, the factor risk premiums for the other three factors are the mean returns of the specific portfolios held long (e.g., the portfolio of small-cap stocks for the "small minus big" factor) minus the mean return for a related but opposite portfolio (e.g., a portfolio of large-cap stocks, in the case of that factor). A portfolio with a sensitivity of 1 to factor *j* and a sensitivity of 0 to all other factors is called a **pure factor portfolio** for factor *j* (or simply the **factor portfolio** for factor *j*).

For example, suppose we have a portfolio with a sensitivity of 1 with respect to Factor 1 and a sensitivity of 0 to all other factors. Using Equation 2, the expected return on this portfolio is $E_1 = R_F + \lambda_1 \times 1$. If $E_1 = 0.12$ and $R_F = 0.04$, then the risk premium for Factor 1 is

$$0.12 = 0.04 + \lambda_1 \times 1.$$

$$\lambda_1 = 0.12 - 0.04 = 0.08, \text{ or } 8\%.$$

EXAMPLE 1

Determining the Parameters in a One-Factor APT Model

Suppose we have three well-diversified portfolios that are each sensitive to the same single factor. Exhibit 1 shows the expected returns and factor sensitivities of these portfolios. Assume that the expected returns reflect a one-year investment horizon. To keep the analysis simple, all investors are assumed to agree upon the expected returns of the three portfolios as shown in the exhibit.

Exhibit 1: Sample Portfolios for a One-Factor Model

Portfolio	Expected Return	Factor Sensitivity
A	0.075	0.5
B	0.150	2.0
C	0.070	0.4

We can use these data to determine the parameters of the APT equation. According to Equation 2, for any well-diversified portfolio and assuming a single factor explains returns, we have $E(R_p) = R_F + \lambda_1\beta_{p,1}$. The factor sensitivities and expected returns are known; thus there are two unknowns, the parameters R_F and λ_1. Because two points define a straight line, we need to set up only two equations. Selecting Portfolios A and B, we have

$$E(R_A) = 0.075 = R_F + 0.5\lambda_1$$

and

$$E(R_B) = 0.150 = R_F + 2\lambda_1.$$

From the equation for Portfolio A, we have $R_F = 0.075 - 0.5\lambda_1$. Substituting this expression for the risk-free rate into the equation for Portfolio B gives

$$0.15 = 0.075 - 0.5\lambda_1 + 2\lambda_1.$$

$$0.15 = 0.075 + 1.5\lambda_1.$$

So, we have $\lambda_1 = (0.15 - 0.075)/1.5 = 0.05$. Substituting this value for λ_1 back into the equation for the expected return to Portfolio A yields

$$0.075 = R_F + 0.05 \times 0.5.$$

$$R_F = 0.05.$$

So, the risk-free rate is 0.05 or 5%, and the factor premium for the common factor is also 0.05 or 5%. The APT equation is

$$E(R_p) = 0.05 + 0.05\beta_{p,1}.$$

From Exhibit 1, Portfolio C has a factor sensitivity of 0.4. Therefore, according to the APT, the expected return of Portfolio C should be

$$E(R_C) = 0.05 + (0.05 \times 0.4) = 0.07,$$

which is consistent with the expected return for Portfolio C given in Exhibit 1.

EXAMPLE 2

Checking Whether Portfolio Returns Are Consistent with No Arbitrage

In this example, we examine how to tell whether expected returns and factor sensitivities for a set of well-diversified portfolios may indicate the presence of an arbitrage opportunity. Exhibit 2 provides data on four hypothetical portfolios. The data for Portfolios A, B, and C are repeated from Exhibit 1. Portfolio D is a new portfolio. The factor sensitivities given relate to the one-factor APT model $E(R_p) = 0.05 + 0.05\beta_{p,1}$ derived in Example 1. As in Example 1, all investors are assumed to agree upon the expected returns of the portfolios. The question raised by the addition of this new Portfolio D is whether the addition of this portfolio created an arbitrage opportunity. If a portfolio can be formed from Portfolios A, B, and C that has the same factor sensitivity as Portfolio D but a different expected return, then an arbitrage opportunity exists: Portfolio D would be either undervalued (if it offers a relatively high expected return) or overvalued (if it offers a relatively low expected return).

Exhibit 2: Sample Portfolios for a One-Factor Model

Portfolio	Expected Return	Factor Sensitivity
A	0.0750	0.50
B	0.1500	2.00
C	0.0700	0.40
D	0.0800	0.45
0.5A + 0.5C	0.0725	0.45

Exhibit 2 gives data for an equally weighted portfolio of A and C. The expected return and factor sensitivity of this new portfolio are calculated as weighted averages of the expected returns and factor sensitivities of A and C. Expected return is thus (0.50)(0.0750) + (0.50)(0.07) = 0.0725, or 7.25%. The factor sensitivity is (0.50)(0.50) + (0.50)(0.40) = 0.45. Note that the factor sensitivity of 0.45 matches the factor sensitivity of Portfolio D. In this case, the configuration of expected returns in relation to factor risk presents an arbitrage opportunity involving Portfolios A, C, and D. Portfolio D offers, at 8%, an expected return that is too high given its factor sensitivity. According to the assumed APT model, the expected return on Portfolio D should be $E(R_D) = 0.05 + 0.05\beta_{D,1} = 0.05 + (0.05 \times 0.45) = 0.0725$, or 7.25%. Portfolio D is undervalued relative to its factor risk. We will buy D (hold it long) in the portfolio that exploits the arbitrage opportunity (the **arbitrage portfolio**). We purchase D using the proceeds from selling short an equally weighted portfolio of A and C with exactly the same 0.45 factor sensitivity as D.

The arbitrage thus involves the following strategy: Invest $10,000 in Portfolio D and fund that investment by selling short an equally weighted portfolio of Portfolios A and C; then close out the investment position at the end of one year (the investment horizon for expected returns). Exhibit 3 demonstrates the arbitrage profits to the arbitrage strategy. The final row of the exhibit shows the net cash flow to the arbitrage portfolio.

Exhibit 3: Arbitrage Opportunity within Sample Portfolios

	Initial Cash Flow	Final Cash Flow	Factor Sensitivity
Portfolio D	−$10,000.00	$10,800.00	0.45
Portfolios A and C	$10,000.00	−$10,725.00	−0.45
Sum	$0.00	$75.00	0.00

As Exhibit 3 shows, if we buy $10,000 of Portfolio D and sell $10,000 of an equally weighted portfolio of Portfolios A and C, we have an initial net cash flow of $0. The expected value of our investment in Portfolio D at the end of one year is $10,000(1 + 0.08) = $10,800. The expected value of our short position in Portfolios A and C at the end of one year is −$10,000(1.0725) = −$10,725. So, the combined expected cash flow from our investment position in one year is $75.

What about the risk? Exhibit 3 shows that the factor risk has been eliminated: Purchasing D and selling short an equally weighted portfolio of A and C creates a portfolio with a factor sensitivity of 0.45 − 0.45 = 0. The portfolios are well diversified, and we assume any asset-specific risk is negligible.

Because an arbitrage is possible, Portfolios A, C, and D cannot all be consistent with the same equilibrium. If Portfolio D actually had an expected return of 8%, investors would bid up its price until the expected return fell and the arbitrage opportunity vanished. Thus, arbitrage restores equilibrium relationships among expected returns.

The Carhart four-factor model, also known as the four-factor model or simply the Carhart model, is a frequently referenced multifactor model in current equity portfolio management practice. Presented in Carhart (1997), it is an extension of the three-factor model developed by Fama and French (1992) to include a momentum factor. According to the model, three groups of stocks tend to have higher returns than those predicted solely by their sensitivity to the market return:

- Small-capitalization stocks
- Low price-to-book stocks, commonly referred to as "value" stocks
- Stocks whose prices have been rising, commonly referred to as "momentum" stocks

On the basis of that evidence, the Carhart model posits the existence of three systematic risk factors beyond the market risk factor. They are named, in the same order as above, the following:

- Small minus big (SMB)
- High minus low (HML)
- Winners minus losers (WML)

Equation 3 is the Carhart model, in which the excess return on the portfolio is explained as a function of the portfolio's sensitivity to a market index (RMRF), a market capitalization factor (SMB), a book-to-market factor (HML), which is essentially the reciprocal of the aforementioned price-to-book ratio, and a momentum factor (WML).

$$R_p - R_F = a_p + b_{p1}\text{RMRF} + b_{p2}\text{SMB} + b_{p3}\text{HML} + b_{p4}\text{WML} + \varepsilon_p, \quad (3)$$

where

R_p and R_F = the return on the portfolio and the risk-free rate of return, respectively

a_p = "alpha" or return in excess of that expected given the portfolio's level of systematic risk (assuming the four factors capture all systematic risk)

b_p = the sensitivity of the portfolio to the given factor

RMRF = the return on a value-weighted equity index in excess of the one-month T-bill rate

SMB = small minus big, a size (market capitalization) factor; SMB is the average return on three small-cap portfolios minus the average return on three large-cap portfolios

HML = high minus low, the average return on two high book-to-market portfolios minus the average return on two low book-to-market portfolios

WML = winners minus losers, a momentum factor; WML is the return on a portfolio of the past year's winners minus the return on a portfolio of the past year's losers. (Note that WML is an equally weighted average of the stocks with the highest 30% 11-month returns lagged 1 month minus the equally weighted average of the stocks with the lowest 30% 11-month returns lagged 1 month.)

ε_p = an error term that represents the portion of the return to the portfolio, p, not explained by the model

Following Equation 2, the Carhart model can be stated as giving equilibrium expected return as

$$E(R_p) = R_F + \beta_{p,1}\text{RMRF} + \beta_{p,2}\text{SMB} + \beta_{p,3}\text{HML} + \beta_{p,4}\text{WML} \quad (4)$$

because the expected value of alpha is zero.

The Carhart model can be viewed as a multifactor extension of the CAPM that explicitly incorporates drivers of differences in expected returns among assets variables that are viewed as anomalies from a pure CAPM perspective. (The term "anomaly" in this context refers to an observed capital market regularity that is not explained by, or contradicts, a theory of asset pricing.) From the perspective of the CAPM, there are size, value, and momentum anomalies. From the perspective of the Carhart model, however, size, value, and momentum represent systematic risk factors; exposure to them is expected to be compensated in the marketplace in the form of differences in mean return.

Size, value, and momentum are common themes in equity portfolio construction, and all three factors continue to have robust uses in active management risk decomposition and return attribution.

TYPES OF MULTIFACTOR MODELS

3

- ☐ describe and compare macroeconomic factor models, fundamental factor models, and statistical factor models
- ☐ describe uses of multifactor models and interpret the output of analyses based on multifactor models

Having introduced the APT, it is appropriate to examine the diversity of multifactor models in current use.

In the following sections, we explain the basic principles of multifactor models and discuss various types of models and their application. We also expand on the APT, which relates the expected return of investments to their risk with respect to a set of factors.

Factors and Types of Multifactor Models

Many varieties of multifactor models have been proposed and researched. We can categorize most of them into three main groups according to the type of factor used:

- In a **macroeconomic factor model**, the factors are surprises in macroeconomic variables that significantly explain returns. In the example of equities, the factors can be understood as affecting either the expected future cash flows of companies or the interest rate used to discount these cash flows back to the present. Among macroeconomic factors that have been used are interest rates, inflation risk, business cycle risk, and credit spreads.
- In a **fundamental factor model**, the factors are attributes of stocks or companies that are important in explaining cross-sectional differences in stock prices. Among the fundamental factors that have been used are the book-value-to-price ratio, market capitalization, the price-to-earnings ratio, and financial leverage.
- In a **statistical factor model**, statistical methods are applied to historical returns of a group of securities to extract factors that can explain the observed returns of securities in the group. In statistical factor models, the factors are actually portfolios of the securities in the group under study and are therefore defined by portfolio weights. Two major types of factor models are factor analysis models and principal components models. In factor analysis models, the factors are the portfolios of securities that best explain (reproduce) historical *return covariances*. In principal components models, the factors are portfolios of securities that best explain (reproduce) the historical *return variances*.

A potential advantage of statistical factor models is that they make minimal assumptions. But the interpretation of statistical factors is generally difficult in contrast to macroeconomic and fundamental factors. A statistical factor that is a portfolio with weights that are similar to market index weights might be interpreted as "the market factor," for example. But in general, associating a statistical factor with economic meaning may not be possible. Because understanding statistical factor models requires substantial preparation in quantitative methods, a detailed discussion of statistical factor models is outside the scope of our coverage.

Our discussion concentrates on macroeconomic factor models and fundamental factor models. Industry use has generally favored fundamental and macroeconomic models, perhaps because such models are much more easily interpreted and rely less on data-mining approaches. Nevertheless, statistical factor models have proponents and are also used in practical applications.

The Structure of Fundamental Factor Models

We earlier gave the equation of a macroeconomic factor model as

$$R_i = a_i + b_{i1}F_1 + b_{i2}F_2 + \ldots + b_{iK}F_K + \varepsilon_i.$$

We can also represent the structure of fundamental factor models with this equation, but we need to interpret the terms differently.

In fundamental factor models, the factors are stated as *returns* rather than return *surprises* in relation to predicted values, so they do not generally have expected values of zero. This approach changes the meaning of the intercept, which is no longer interpreted as the expected return. Note that if the coefficients were not standardized, as described in the following paragraph, the intercept could be interpreted as the risk-free rate because it would be the return to an asset with no factor risk (zero factor betas) and no asset-specific risk (with standardized coefficients, the intercept is not interpreted beyond being an intercept in a regression included so that the expected asset-specific risk equals zero).

Factor sensitivities are also interpreted differently in most fundamental factor models. In fundamental factor models, the factor sensitivities are attributes of the security. An asset's sensitivity to a factor is expressed using a **standardized beta**: the value of the attribute for the asset minus the average value of the attribute across all stocks divided by the standard deviation of the attribute's values across all stocks.

$$b_{ik} = \frac{\text{Value of attribute } k \text{ for asset } i - \text{Average value of attribute } k}{\sigma(\text{Values of attribute } k)}. \quad (5)$$

Consider a fundamental model for equities that uses a dividend yield factor. After standardization, a stock with an average dividend yield will have a factor sensitivity of 0; a stock with a dividend yield one standard deviation above the average will have a factor sensitivity of 1; and a stock with a dividend yield one standard deviation below the average will have a factor sensitivity of −1. Suppose, for example, that an investment has a dividend yield of 3.5% and that the average dividend yield across all stocks being considered is 2.5%. Further, suppose that the standard deviation of dividend yields across all stocks is 2%. The investment's sensitivity to dividend yield is (3.5% − 2.5%)/2% = 0.50, or one-half standard deviation above average. The scaling permits all factor sensitivities to be interpreted similarly, despite differences in units of measure and scale in the variables. The exception to this interpretation is factors for binary variables, such as industry membership. A company either participates in an industry or does not. The industry factor is represented by dummy variables: The value of the variable is 1 if the stock belongs to the industry and 0 if it does not.

A second distinction between macroeconomic multifactor models and fundamental factor models is that with the former, we develop the factor (surprise) series first and then estimate the factor sensitivities through regressions. With the latter, we generally specify the factor sensitivities (attributes) first and then estimate the factor returns through regressions.

Financial analysts use fundamental factor models for a variety of purposes, including portfolio performance attribution and risk analysis. (*Performance attribution* consists of return attribution and risk attribution. *Return attribution* is a set of techniques used to identify the sources of the excess return of a portfolio against its benchmark. *Risk attribution* addresses the sources of risk, identifying the sources of portfolio volatility for absolute mandates and the sources of tracking risk for relative mandates.) Fundamental factor models focus on explaining the returns to individual stocks using observable fundamental factors that describe either attributes of the securities themselves or attributes of the securities' issuers. Industry membership, price-to-earnings ratio, book-value-to-price ratio, size, and financial leverage are examples of fundamental factors.

Example 4 discusses a study that examined macroeconomic, fundamental, and statistical factor models.

We encounter a range of distinct representations of risk in the fundamental models that are currently used in practical applications. Diversity exists in both the identity and exact definition of factors as well as in the underlying functional form and estimation procedures. Despite the diversity, we can place the factors of most fundamental factor models for equities into three broad groups:

- **Company fundamental factors**. These are factors related to the company's internal performance. Examples are factors relating to earnings growth, earnings variability, earnings momentum, and financial leverage.
- **Company share-related factors**. These factors include valuation measures and other factors related to share price or the trading characteristics of the shares. In contrast to the previous category, these factors directly incorporate investors' expectations concerning the company. Examples include price multiples, such as earnings yield, dividend yield, and book to market. Market capitalization falls under this heading. Various models incorporate variables relating to share price momentum, share price volatility, and trading activity that fall in this category.
- **Macroeconomic factors**. Sector or industry membership factors fall under this heading. Various models include such factors as CAPM beta, other similar measures of systematic risk, and yield curve level sensitivity—all of which can be placed in this category.

For global factor models, in particular, a classification of country, industry, and style factors is often used. In that classification, country and industry factors are dummy variables for country and industry membership, respectively. Style factors include those related to earnings, risk, and valuation that define types of securities typical of various styles of investing.

Fixed-Income Multifactor Models

While the previous discussion focuses on equity applications, similar approaches are equally suited to fixed income. In addition, some of the same broad factor groupings are relevant for bonds.

Macroeconomic Multifactor Models

Macroeconomic models, as discussed earlier, are easily translatable to fixed-income investing. For instance, surprises to economic growth, interest rates, and inflation will impact bond pricing, often mechanically.

Consider a bond factor model in which the returns are correlated with two factors. Following our earlier discussion, returns for bonds are assumed to be correlated with surprises in inflation rates and surprises in GDP growth. The return to *bond i*, R_i, can be modeled as

$$R_i = a_i + b_{i1}F_{INFL} + b_{i2}F_{GDP} + \varepsilon_i,$$

where

R_i = the return to bond i

a_i = the expected return to bond i

b_{i1} = the sensitivity of the return on bond i to inflation rate surprises

F_{INFL} = the surprise in inflation rates

b_{i2} = the sensitivity of the return on bond i to GDP growth surprises

F_{GDP} = the surprise in GDP growth (assumed to be uncorrelated with F_{INFL})

ε_i = an error term with a zero mean that represents the portion of the return to bond i not explained by the factor model

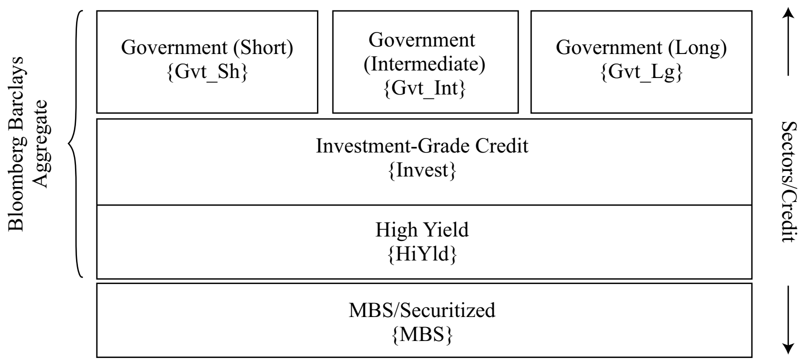
Fundamental Multifactor Models

Fundamental factor approaches have been developed to address the unique aspects of fixed income by using, for example, the following categories:

- Duration (ranging from cash to long-dated bonds)
- Credit (ranging from government securities to high yield)
- Currency (ranging from home currency to foreign developed and emerging market currencies)
- Geography (specific developed and emerging markets)

A simplified structure, shown in Exhibit 4, divides the US Barclays Bloomberg Aggregate index, a standard bond benchmark, into sectors, where each has such unique factor exposures as spread or duration. This factor model was developed by Dopfel (2004), and the factors have been chosen to cover three macro sectors plus high yield. The government sector is further broken down into three maturity buckets to help explain duration exposures.

Exhibit 4: A Simple Fixed-Income Fundamental Framework



Source: Dopfel (2004).

These components can be thought of as both macroeconomic and fundamental. They are macroeconomically oriented because spread, or expected return above similar duration government bonds, is closely related to the growth factor and is sometimes expressed as simply credit spread. Fundamentally, duration can also be thought of as a factor. This simplistic approach can be extended to encompass global fixed-income markets or adapted to a specific country's market:

$$R_i = a_i + b_{i1}F_{Gvt\_Sh} + b_{i2}F_{Gvt\_Int} + b_{i3}F_{Gvt\_Lg} + b_{i4}F_{Invest} + b_{i5}F_{HiYld} + b_{i6}F_{MBS} + \varepsilon_i,\text{where}$$

R_i = the return to bond i

a_i = the expected return to bond i

b_{ik} = the sensitivity of the return on bond i to factor k

F_k = factor k, where k represents "Gov't (Short)," "Gov't (Long)," and so on

ε_i = an error term with a zero mean that represents the portion of the return to bond i not explained by the factor model

The historic style factor weights, b_{ik}, are determined by a constrained regression (the constraint being that the total "weights" add up to 100%) of the portfolio returns against the listed style factors.

This framework lends itself readily to performance and risk attribution, along with portfolio construction. When evaluating a fixed-income manager, such characteristics as spread, duration, yield, and quality can be incorporated. This type of framework can also be extended to ESG (environmental, social, and governance) considerations as these should be generally unrelated to the basic duration and spread foundation presented. For instance, each box in Exhibit 4 could also contain E, S, and G scores, which after the initial disaggregation of a fixed-income return stream into duration and spread components could be used to model the overall portfolio's aggregate scores. For forward-looking portfolio construction purposes, a desired loading on duration, spread, and ESG scores could be handled with a quantitative objective function.

Risk and Style Multifactor Models

Another category of multifactor approach incorporates risk, or style, factors, several of which can thematically apply across asset classes. Examples of such factors include momentum, value, carry, and volatility. Many of these are similar in construction to those commonly used in equity portfolios. Examples include defining value as real (inflation-adjusted) yield, momentum as the previous 12-month excess return, and carry as the term spread. An illustrative example of risk factor approaches, in this case across asset classes, can be found in Exhibit 5.

Exhibit 5: An Illustration of Factor Approaches across Asset Classes

	Factor/Asset Class	Equity	Credit	Treasury	Commodities	Currency
Macro	Economic Growth	xx	x			
	Rates		x	xx		
	Inflation			x	xx	x

	Factor/Asset Class	Equity	Credit	Treasury	Commodities	Currency
Style	Value	xx	x		x	x
	Size	xx				
	Momentum	xx	xx	xx	xx	xx
	Carry	x	xx	xx	xx	xx
	Low-Volume	xx	x			

Note: Double check marks denote strong alignment between risk factor and asset class; single check marks denote moderate alignment.
*Source:*Podkaminer (2017).

Of the three types of multifactor models (macroeconomic, fundamental, and statistical), statistical models can be most easily applied to various asset classes, including fixed income, as no asset-class-specific tuning is required given the minimal required assumption set. This is in contrast to macroeconomic and fundamental models, which both require adjustments and repurposing to ensure the frameworks are fit for the specifics of bond investing. Example 3 shows how expected return could be expressed.

EXAMPLE 3

Calculating Factor-Based Expected Returns at the Portfolio Level

1. A fixed-income portfolio has the following estimated exposures: 35% intermediate government bonds, 40% investment-grade credit, 5% securitized, and 20% high yield. The expected component returns are

   A. Short government bonds: 0.25%
   B. Intermediate government bonds: 1.50%
   C. Long government bonds: 3.00%
   D. Investment-grade credit: 4.25%
   E. MBS/Securitized: 1.75%
   F. High yield: 5.75%
   G. Express the expected return of the portfolio.

Solution

Expected return could be expressed as

$E(R) = 3.46\%$

$= (0.35)(1.50\%) + (0.40)(4.25\%) + (0.05)(1.75\%) + (0.20)(5.75\%)$.

EXAMPLE 4

Reconciling Bond Portfolio Characteristics Using Style Factors

Talia Ayalon is evaluating intermediate duration (between 5 and 7 years) investment-grade fixed-income strategies using the framework presented in Exhibit 4. One of the strategies has the following sector attribution (totaling to 100%):

Gov't (Short) 2%	Gov't (Intermediate) 4%	Gov't (Long) 14%
Investment-Grade Credit 56%		
MBS/Securitized 6%		
High Yield 18%		

Are these sector exposures consistent with an intermediate duration investment-grade approach? Why or why not?

Suggested answer:

No, the sector exposures are inconsistent with the stated approach for two reasons: 1) The 18% exposure to high yield constitutes a significant amount of below investment-grade exposure. A true investment-grade portfolio would, for example, not have exposure to high yield. 2) The loading to longer duration sectors implies a longer-than-intermediate duration for the portfolio.

MACROECONOMIC FACTOR MODELS

4

- ☐ calculate the expected return on an asset given an asset's factor sensitivities and the factor risk premiums
- ☐ describe and compare macroeconomic factor models, fundamental factor models, and statistical factor models

The representation of returns in macroeconomic factor models assumes that the returns to each asset are correlated with only the surprises in some factors related to the aggregate economy, such as inflation or real output. We can define *surprise* in general as the actual value minus predicted (or expected) value. A factor's surprise is the component of the factor's return that was unexpected, and the factor surprises constitute the model's independent variables. This idea contrasts with the representation of independent variables as returns in Equation 2, reflecting the fact that how the independent variables are represented varies across different types of models.

Suppose that K macro factors explain asset returns. Then in a macroeconomic factor model, Equation 6 expresses the return of asset i:

$$R_i = a_i + b_{i1}F_1 + b_{i2}F_2 + \ldots + b_{iK}F_K + \varepsilon_i, \quad (6)$$

where

R_i = the return to asset i

a_i = the expected return to asset i

b_{ik} = the sensitivity of the return on asset i to a surprise in factor k, $k = 1, 2, \ldots, K$

F_k = the surprise in the factor k, $k = 1, 2, \ldots, K$

ε_i = an error term with a zero mean that represents the portion of the return to asset i not explained by the factor model

Surprise in a macroeconomic factor can be illustrated as follows: Suppose we are analyzing monthly returns for stocks. At the beginning of each month, we have a prediction of inflation for the month. The prediction may come from an econometric model or a professional economic forecaster, for example. Suppose our forecast at the beginning of the month is that inflation will be 0.4% during the month. At the end of the month, we find that inflation was actually 0.5% during the month. During any month,

Actual inflation = Predicted inflation + Surprise inflation.

In this case, actual inflation was 0.5% and predicted inflation was 0.4%. Therefore, the surprise in inflation was 0.5% − 0.4% = 0.1%.

What is the effect of defining the factors in terms of surprises? Suppose we believe that inflation and gross domestic product (GDP) growth are two factors that carry risk premiums; that is, inflation and GDP represent priced risk. (GDP is a money measure of the goods and services produced within a country's borders.) We do not use the predicted values of these variables because the predicted values should already be reflected in stock prices and thus in their expected returns. The intercept a_i, the expected return to asset i, reflects the effect of the predicted values of the macroeconomic variables on expected stock returns. The surprise in the macroeconomic variables during the month, however, contains new information about the variable. As a result, this model structure analyzes the return to an asset in three components: the asset's expected return, its unexpected return resulting from new information about the factors, and an error term.

Consider a factor model in which the returns to each asset are correlated with two factors. For example, we might assume that the returns for a particular stock are correlated with surprises in inflation rates and surprises in GDP growth. For stock i, the return to the stock can be modeled as

$$R_i = a_i + b_{i1}F_{INFL} + b_{i2}F_{GDP} + \varepsilon_i,$$

where

R_i = the return to stock i

a_i = the expected return to stock i

b_{i1} = the sensitivity of the return on stock i to inflation rate surprises

F_{INFL} = the surprise in inflation rates

b_{i2} = the sensitivity of the return on stock i to GDP growth surprises

F_{GDP} = the surprise in GDP growth (assumed to be uncorrelated with F_{INFL})

ε_i = an error term with a zero mean that represents the portion of the return to asset i not explained by the factor model

Consider first how to interpret b_{i1}. The factor model predicts that a 1 percentage point surprise in inflation rates will contribute b_{i1} percentage points to the return to stock i. The slope coefficient b_{i2} has a similar interpretation relative to the GDP growth factor. Thus, slope coefficients are naturally interpreted as the factor sensitivities of the asset. A *factor sensitivity* is a measure of the response of return to each unit of increase in a factor, holding all other factors constant. (Factor sensitivities are sometimes called *factor betas* or *factor loadings*.)

Now consider how to interpret the intercept a_i. Recall that the error term has a mean or average value of zero. If the surprises in both inflation rates and GDP growth are zero, the factor model predicts that the return to asset i will be a_i. Thus, a_i is the expected value of the return to stock i.

Finally, consider the error term, ε_i. The intercept a_i represents the asset's expected return. The term ($b_{i1}F_{INFL} + b_{i2}F_{GDP}$) represents the return resulting from factor surprises, and we have interpreted these as the sources of risk shared with other assets. The term ε_i is the part of return that is unexplained by expected return or the factor surprises. If we have adequately represented the sources of common risk (the factors), then ε_i must represent an asset-specific risk. For a stock, it might represent the return from an unanticipated company-specific event.

The risk premium for the GDP growth factor is typically positive. The risk premium for the inflation factor, however, is typically negative. Thus, an asset with a positive sensitivity to the inflation factor—an asset with returns that tend to be positive in response to unexpectedly high inflation—would have a lower required return than if its inflation sensitivity were negative; an asset with positive sensitivity to inflation would be in demand for its inflation-hedging ability.

This discussion has broader applications. It can be used for various asset classes, including fixed income and commodities. It can also be used in asset allocation, where asset classes can be examined in relation to inflation and GDP growth, as illustrated in the following exhibit. In Exhibit 6, each quadrant reflects a unique mix of inflation and economic growth expectations. Certain asset classes or securities can be expected to perform differently in various inflation and GDP growth regimes and can be plotted in the appropriate quadrant, thus forming a concrete illustration of a two-factor model.

Exhibit 6: Growth and Inflation Factor Matrix

	Inflation	
Growth	**Low Inflation/Low Growth** ▪ Cash ▪ Government bonds	**High Inflation/Low Growth** ▪ Inflation-linked bonds ▪ Commodities ▪ Infrastructure
	Low Inflation/High Growth ▪ Equity ▪ Corporate debt	**High Inflation/High Growth** ▪ Real assets (real estate, timberland, farmland, energy)

Note: Entries are assets likely to benefit from the specified combination of growth and inflation.

In macroeconomic factor models, the time series of factor surprises are constructed first. Regression analysis is then used to estimate assets' sensitivities to the factors. In practice, estimated sensitivities and intercepts are often acquired from one of the many consulting companies that specialize in factor models. When we have the parameters for the individual assets in a portfolio, we can calculate the portfolio's parameters as a weighted average of the parameters of individual assets. An individual asset's weight in that calculation is the proportion of the total market value of the portfolio that the individual asset represents.

EXAMPLE 5

Estimating Returns for a Two-Stock Portfolio Given Factor Sensitivities

Suppose that stock returns are affected by two common factors: surprises in inflation and surprises in GDP growth. A portfolio manager is analyzing the returns on a portfolio of two stocks, Manumatic (MANM) and Nextech (NXT). The following equations describe the returns for those stocks, where the factors F_{INFL} and F_{GDP} represent the surprise in inflation and GDP growth, respectively:

$$R_{MANM} = 0.09 - 1F_{INFL} + 1F_{GDP} + \varepsilon_{MANM}.$$

$$R_{NXT} = 0.12 + 2F_{INFL} + 4F_{GDP} + \varepsilon_{NXT}.$$

One-third of the portfolio is invested in Manumatic stock, and two-thirds is invested in Nextech stock.

In evaluating the equations for surprises in inflation and GDP, convert amounts stated in percentage terms to decimal form.

1. Formulate an expression for the return on the portfolio.

Solution to 1:

The portfolio's return is the following weighted average of the returns to the two stocks:

$$R_P = (1/3)(0.09) + (2/3)(0.12) + [(1/3)(-1) + (2/3)(2)]F_{INFL} + [(1/3)(1) + (2/3)(4)]F_{GDP} + (1/3)\varepsilon_{MANM} + (2/3)\varepsilon_{NXT}$$

$$= 0.11 + 1F_{INFL} + 3F_{GDP} + (1/3)\varepsilon_{MANM} + (2/3)\varepsilon_{NXT}.$$

2. State the expected return on the portfolio.

Solution to 2:

The expected return on the portfolio is 11%, the value of the intercept in the expression obtained in the solution to 1.

3. Calculate the return on the portfolio given that the surprises in inflation and GDP growth are 1% and 0%, respectively, assuming that the error terms for MANM and NXT both equal 0.5%.

Solution to 3:

$$R_P = 0.11 + 1F_{INFL} + 3F_{GDP} + (1/3)\varepsilon_{MANM} + (2/3)\varepsilon_{NXT}$$

$$= 0.11 + 1(0.01) + 3(0) + (1/3)(0.005) + (2/3)(0.005)$$

$$= 0.125, \text{ or } 12.5\%.$$

FUNDAMENTAL FACTOR MODELS

5

- [ ] calculate the expected return on an asset given an asset's factor sensitivities and the factor risk premiums
- [ ] describe and compare macroeconomic factor models, fundamental factor models, and statistical factor models

EXAMPLE 6

Comparing Types of Factor Models

Connor (1995) contrasted a macroeconomic factor model with a fundamental factor model to compare how well the models explain stock returns.

Connor reported the results of applying a macroeconomic factor model to the returns for 779 large-cap US stocks based on monthly data from January 1985 through December 1993. Using five macroeconomic factors, Connor was able to explain approximately 11% of the variance of return on these stocks. Exhibit 7 shows his results.

Exhibit 7: The Explanatory Power of the Macroeconomic Factors

Factor	Explanatory Power from Using Each Factor Alone	Increase in Explanatory Power from Adding Each Factor to All the Others
Inflation	1.3%	0.0%
Term structure	1.1%	7.7%
Industrial production	0.5%	0.3%
Default premium	2.4%	8.1%
Unemployment	–0.3%	0.1%
All factors (total explanatory power)		10.9%

Notes: The explanatory power of a given model was computed as 1 – [(Average asset –Specific variance of return across stocks)/(Average total variance of return across stocks)]. The variance estimates were corrected for degrees of freedom, so the marginal contribution of a factor to explanatory power can be zero or negative. Explanatory power captures the proportion of the total variance of return that a given model explains for the average stock.

Source: Connor (1995).

Connor also reported a fundamental factor analysis of the same companies. The factor model employed was the BARRA US-E2 model (as of 2019, the current version is E4). Exhibit 8 shows these results. In the exhibit, "variability in markets" represents the stock's volatility, "success" is a price momentum variable, "trade activity" distinguishes stocks by how often their shares trade, and "growth" distinguishes stocks by past and anticipated earnings growth (explanations of variables are from Grinold and Kahn 1994).

Exhibit 8: The Explanatory Power of the Fundamental Factors

Factor	Explanatory Power from Using Each Factor Alone	Increase in Explanatory Power from Adding Each Factor to All the Others
Industries	16.3%	18.0%
Variability in markets	4.3%	0.9%
Success	2.8%	0.8%
Size	1.4%	0.6%
Trade activity	1.4%	0.5%
Growth	3.0%	0.4%
Earnings to price	2.2%	0.6%
Book to price	1.5%	0.6%
Earnings variability	2.5%	0.4%
Financial leverage	0.9%	0.5%
Foreign investment	0.7%	0.4%
Labor intensity	2.2%	0.5%
Dividend yield	2.9%	0.4%
All factors (total explanatory power)		42.6%

Source: Connor (1995).

As Exhibit 8 shows, the most important fundamental factor is "industries," represented by 55 industry dummy variables. The fundamental factor model explained approximately 43% of the variation in stock returns, compared with approximately 11% for the macroeconomic factor model. Because "industries" must sum to the market and the market portfolio is not incorporated in the macroeconomic factor model, some advantage to the explanatory power of the fundamental factor may be built into the specific models being compared. Connor's article also does not provide tests of the statistical significance of the various factors in either model; however, Connor's research is strong evidence for the usefulness of fundamental factor models. Moreover, this evidence is mirrored by the wide use of those models in the investment community. For example, fundamental factor models are frequently used in portfolio performance attribution. Typically, fundamental factor models employ many more factors than macroeconomic factor models, giving a more detailed picture of the sources of an investment manager's returns.

We cannot conclude from this study, however, that fundamental factor models are inherently superior to macroeconomic factor models. Each major type of model has its uses. The factors in various macroeconomic factor models are individually backed by statistical evidence that they represent systematic risk (i.e., risk that cannot be diversified away). The same may not be true of each factor in a fundamental factor model. For example, a portfolio manager can easily construct a portfolio that excludes a particular industry, so exposure to a particular industry is not systematic risk.

The two types of factors, macroeconomic and fundamental, have different implications for measuring and managing risk, in general. The macroeconomic factor set is parsimonious (five variables in the model studied) and allows a portfolio manager to incorporate economic views into portfolio construction by adjustments to portfolio exposures to macro factors. The fundamental factor set examined by Connor is large (67 variables, including the 55 industry dummy variables); at the expense of greater complexity, it can give a more detailed picture of risk in terms that are easily related to company and security characteristics. Connor found that the macroeconomic factor model had no marginal explanatory power when added to the fundamental factor model, implying that the fundamental risk attributes capture all the risk characteristics represented by the macroeconomic factor betas. Because the fundamental factors supply such a detailed description of the characteristics of a stock and its issuer, however, this finding is not necessarily surprising.

FACTOR MODELS IN RETURN ATTRIBUTION

6

- ☐ calculate the expected return on an asset given an asset's factor sensitivities and the factor risk premiums
- ☐ describe uses of multifactor models and interpret the output of analyses based on multifactor models
- ☐ describe the potential benefits for investors in considering multiple risk dimensions when modeling asset returns

The following sections present selected applications of multifactor models in investment practice. The applications discussed are return attribution, risk attribution, portfolio construction, and strategic portfolio decisions. We begin by discussing portfolio return attribution and risk attribution, focusing on the analysis of benchmark-relative returns. After discussing performance attribution and risk analysis, we explain the use of multifactor models in creating a portfolio with a desired set of risk exposures.

Additionally, multifactor models can be used for asset allocation purposes. Some large, sophisticated asset owners have chosen to define their asset allocation opportunity sets in terms of macroeconomic or thematic factors and aggregate factor exposures (represented by pure factor portfolios as defined earlier). Many others are examining their traditionally derived asset allocation policies using factor models to map asset class exposure to factor sensitivities. The trend toward factor-based asset allocation has two chief causes: First is the increasing availability of sophisticated factor models (like the BARRA models used in the following examples); second is the more intense focus by asset owners on the many dimensions of risk.

Factor Models in Return Attribution

Multifactor models can help us understand in detail the sources of a manager's returns relative to a benchmark. For simplicity, in this section we analyze the sources of the returns of a portfolio fully invested in the equities of a single national equity market, which allows us to ignore the roles of country selection, asset allocation, market timing, and currency hedging. The same methodology can, however, be applied across asset classes and geographies.

Analysts often favor fundamental multifactor models in decomposing (separating into basic elements) the sources of returns. In contrast to statistical factor models, fundamental factor models allow the sources of portfolio performance to be described using commonly understood terms. Fundamental factors are also thematically understandable and can be incorporated into simple narratives for clients concerning return or risk attribution.

Also, in contrast to macroeconomic factor models, fundamental models express investment style choices and security characteristics more directly and often in greater detail.

We first need to understand the objectives of active managers. As mentioned previously, managers are commonly evaluated relative to a specified benchmark. Active portfolio managers hold securities in different-from-benchmark weights in an attempt to add value to their portfolios relative to a passive investment approach. Securities held in different-from-benchmark weights reflect portfolio manager expectations that differ from consensus expectations. For an equity manager, those expectations may relate to common factors driving equity returns or to considerations unique to a company. Thus, when we evaluate an active manager, we want to ask such questions as, Did the manager have insights that were effectively translated into returns in excess of those that were available from a passive alternative? Analyzing the sources of returns using multifactor models can help answer these questions.

The return on a portfolio, R_p, can be viewed as the sum of the benchmark's return, R_B, and the **active return** (portfolio return minus benchmark return):

$$\text{Active return} = R_p - R_B. \tag{7}$$

With the help of a factor model, we can analyze a portfolio manager's active return as the sum of two components. The first component is the product of the portfolio manager's factor tilts (over- or underweights relative to the benchmark factor sensitivities) and the factor returns; we call this component the return from factor tilts. The second component of active return reflects the manager's skill in individual asset selection (ability to overweight securities that outperform the benchmark or underweight securities that underperform the benchmark); we call this component security selection. Equation 8 shows the decomposition of active return into those two components, where k represents the factor or factors represented in the benchmark portfolio:

$$\begin{aligned}\text{Active return} &= \sum_{k=1}^{K}\left[(\text{Portfolio sensitivity})_k - (\text{Benchmark sensitivity})_k\right] \\ &\times (\text{Factor return})_k + \text{Security selection}\end{aligned} \tag{8}$$

In Equation 8, the portfolio's and benchmark's sensitivities to each factor are calculated as of the beginning of the evaluation period.

EXAMPLE 7

Four-Factor Model Active Return Decomposition

As an equity analyst at a pension fund sponsor, Ronald Service uses the Carhart four-factor multifactor model of Equation 3a to evaluate US equity portfolios:

$$R_p - R_F = a_p + b_{p1}\text{RMRF} + b_{p2}\text{SMB} + b_{p3}\text{HML} + b_{p4}\text{WML} + \varepsilon_p.$$

Service's current task is to evaluate the performance of the most recently hired US equity manager. That manager's benchmark is an index representing the performance of the 1,000 largest US stocks by market value. The manager describes himself as a "stock picker" and points to his performance in beating the benchmark as evidence that he is successful. Exhibit 9 presents an analysis based on the Carhart model of the sources of that manager's active return during the year, given an assumed set of factor returns. In Exhibit 9, the entry "A. Return from Factor Tilts = 2.1241%" is the sum of the four numbers above it. The entry "B. Security Selection" gives security selection as equal to −0.05%. "C. Active Return" is found as the sum of these two components: 2.1241% + (−0.05%) = 2.0741%.

Exhibit 9: Active Return Decomposition

	Factor Sensitivity				Contribution to Active Return	
Factor	Portfolio (1)	Benchmark (2)	Difference (3) = (1) − (2)	Factor Return (4)	Absolute (3) × (4)	Proportion of Total Active
RMRF	0.95	1.00	−0.05	5.52%	−0.2760%	−13.3%
SMB	−1.05	−1.00	−0.05	−3.35%	0.1675%	8.1%
HML	0.40	0.00	0.40	5.10%	2.0400%	98.4%
WML	0.05	0.03	0.02	9.63%	0.1926%	9.3%
				A. Return from Factor Tilts =	2.1241%	102.4%
				B. Security Selection =	−0.0500%	−2.4%
				C. Active Return (A + B) =	2.0741%	100.0%

From his previous work, Service knows that the returns to growth-style portfolios often have a positive sensitivity to the momentum factor (WML). By contrast, the returns to certain value-style portfolios, in particular those following a contrarian strategy, often have a negative sensitivity to the momentum factor. Using the information given, address the following questions (assume the benchmark chosen for the manager is appropriate):

1. Determine the manager's investment mandate and his actual investment style.

Solution:

The benchmarks chosen for the manager should reflect the baseline risk characteristics of the manager's investment opportunity set and his mandate. We can ascertain whether the manager's actual style follows the mandate by examining the portfolio's actual factor exposures:

- The sensitivities of the benchmark are consistent with the description in the text. The sensitivity to RMRF of 1 indicates that the assigned benchmark has average market risk, consistent with it being a broad-based index; the negative sensitivity to SMB indicates a large-cap orientation. The mandate might be described as large-cap without a value/growth bias (HML is zero) or a momentum bias (WML is close to zero).
- Stocks with high book-to-market ratios are generally viewed as value stocks. Because the equity manager has a positive sensitivity to HML (0.40), it appears that the manager has a value orientation. The manager is approximately neutral to the momentum factor, so the equity manager is not a momentum investor and probably not a contrarian value investor. In summary, these considerations suggest that the manager has a large-cap value orientation.

2. Evaluate the sources of the manager's active return for the year.

Solution:

The dominant source of the manager's positive active return was his positive active exposure to the HML factor. The bet contributed approximately 98% of the realized active return of about 2.07%. The manager's active exposure to the overall market (RMRF) was unprofitable, but his active exposures to small stocks (SMB) and to momentum (WML) were profitable. The magnitudes of the manager's active exposures to RMRF, SMB, and WML were relatively small, however, so the effects of those bets on active return were minor compared with his large and successful bet on HML.

3. What concerns might Service discuss with the manager as a result of the return decomposition?

Solution:

Although the manager is a self-described "stock picker," his active return from security selection in this period was actually negative. His positive active return resulted from the concurrence of a large active bet on HML and a high return to that factor during the period. If the market had favored growth rather than value without the manager doing better in individual security selection, the manager's performance would have been unsatisfactory. Service's conversations with the manager should focus on evidence that he can predict changes in returns to the HML factor and on the manager's stock selection discipline.

7 FACTOR MODELS IN RISK ATTRIBUTION

- ☐ explain sources of active risk and interpret tracking risk and the information ratio
- ☐ describe uses of multifactor models and interpret the output of analyses based on multifactor models
- ☐ describe the potential benefits for investors in considering multiple risk dimensions when modeling asset returns

Building on the discussion of active returns, this section explores the analysis of active risk. A few key terms are important to the understanding of how factor models are used to build an understanding of a portfolio manager's risk exposures. We will describe them briefly before moving on to the detailed discussion of risk attribution.

Active risk can be represented by the standard deviation of active returns. A traditional term for that standard deviation is **tracking error** (TE). **Tracking risk** is a synonym for tracking error that is often used in the CFA Program curriculum. We will use the abbreviation TE for the concept of active risk and refer to it usually as tracking error:

$$TE = s(R_p - R_B). \quad (9)$$

In Equation 9, $s(R_p - R_B)$ indicates that we take the sample standard deviation (indicated by s) of the time series of differences between the portfolio return, R_p, and the benchmark return, R_B. We should be careful that active return and tracking error are stated on the same time basis. As an approximation assuming returns are serially uncorrelated, to annualize a daily TE based on daily returns, we multiply daily TE by $(250)^{1/2}$ based on 250 trading days in a year. To annualize a monthly TE based on monthly returns, we multiply monthly TE by $(12)^{1/2}$.

As a broad indication of the range for tracking error, in US equity markets a well-executed passive investment strategy can often achieve a tracking error on the order of 0.10% or less per year. A low-risk active or enhanced index investment strategy, which makes tightly controlled use of managers' expectations, often has a tracking error goal of 2% per year. A diversified active large-cap equity strategy that might be benchmarked to the S&P 500 Index would commonly have a tracking error in the range of 2%–6% per year. An aggressive active equity manager might have a tracking error in the range of 6%–10% or more.

Somewhat analogous to the use of the traditional Sharpe measure in evaluating absolute returns, the **information ratio** (IR) is a tool for evaluating mean active returns per unit of active risk. The historical or *ex post* IR is expressed as follows:

$$\text{IR} = \frac{\overline{R}_p - \overline{R}_B}{s\left(R_p - R_B\right)}. \tag{10}$$

In the numerator of Equation 10, $\overline{R}_p$ and $\overline{R}_B$ stand for the sample mean return on the portfolio and the sample mean return on the benchmark, respectively. The equation assumes that the portfolio being evaluated has the same systematic risk as its benchmark. To illustrate the calculation, if a portfolio achieved a mean return of 9% during the same period that its benchmark earned a mean return of 7.5% and the portfolio's tracking error (the denominator) was 6%, we would calculate an information ratio of (9% − 7.5%)/6% = 0.25. Setting guidelines for acceptable active risk or tracking error is one of the methods that some investors use to ensure that the overall risk and style characteristics of their investments are in line with their chosen benchmark.

Note that in addition to focusing exclusively on *active* risk, multifactor models can also be used to decompose and attribute sources of *total* risk. For instance, a multi-asset class multi-strategy long/short fund can be evaluated with an appropriate multifactor model to reveal insights on sources of total risk.

EXAMPLE 8

Creating Active Manager Guidelines

The framework of active return and active risk is appealing to investors who want to manage the risk of investments. The benchmark serves as a known and continuously observable reference standard in relation to which quantitative risk and return objectives may be stated and communicated. For example, a US public employee retirement system invited investment managers to submit proposals to manage a "low-active-risk US large-cap equity fund" that would be subject to the following constraints:

- Shares must be components of the S&P 500.
- The portfolio should have a minimum of 200 issues. At time of purchase, the maximum amount that may be invested in any one issuer is 5% of the portfolio at market value or 150% of the issuers' weight within the S&P 500, whichever is greater.
- The portfolio must have a minimum information ratio of 0.30 either since inception or over the last seven years.
- The portfolio must also have tracking risk of less than 3% with respect to the S&P 500 either since inception or over the last seven years.

Once a suitable active manager is found and hired, these requirements can be written into the manager's guidelines. The retirement system's individual mandates would be set such that the sum of mandates across managers would equal the desired risk exposures.

Analysts use multifactor models to understand a portfolio manager's risk exposures in detail. By decomposing active risk, the analyst's objective is to measure the portfolio's active exposure along each dimension of risk—in other words, to understand the sources of tracking error. This can even be done at the level of individual holdings. Among the questions analysts will want to answer are the following:

- What active exposures contributed most to the manager's tracking error?
- Was the portfolio manager aware of the nature of his active exposures, and if so, can he articulate a rationale for assuming them?
- Are the portfolio's active risk exposures consistent with the manager's stated investment philosophy?
- Which active bets earned adequate returns for the level of active risk taken?

In addressing these questions, analysts often choose fundamental factor models because they can be used to relate active risk exposures to a manager's portfolio decisions in a fairly direct and intuitive way. In this section, we explain how to decompose or explain a portfolio's active risk using a multifactor model.

We previously addressed the decomposition of active return; now we address the decomposition of active risk. In analyzing risk, it is more convenient to use variances rather than standard deviations because the variances of uncorrelated variables are additive. We refer to the variance of active return as **active risk squared**:

$$\text{Active risk squared} = s^2(R_p - R_B). \qquad (11)$$

We can separate a portfolio's active risk squared into two components:

- **Active factor risk** is the contribution to active risk squared resulting from the portfolio's different-from-benchmark exposures relative to factors specified in the risk model.
- **Active specific risk** or **security selection risk** measures the active non-factor or residual risk assumed by the manager. Portfolio managers attempt to provide a positive average return from security selection as compensation for assuming active specific risk.

As we use the terms, "active specific risk" and "active factor risk" refer to variances rather than standard deviations. When applied to an investment in a single asset class, active risk squared has two components:

$$\text{Active risk squared} = \text{Active factor risk} + \text{Active specific risk}. \qquad (12)$$

Active factor risk represents the part of active risk squared explained by the portfolio's active factor exposures. Active factor risk can be found indirectly as the risk remaining after active specific risk is deducted from active risk squared. Active specific risk can be expressed as

$$\text{Active specific risk} = \sum_{i=1}^{n} \left(w_i^a\right)^2 \sigma_{\varepsilon_i}^2,$$

where w_i^a is the ith asset's active weight in the portfolio (that is, the difference between the asset's weight in the portfolio and its weight in the benchmark) and $\sigma_{\varepsilon_i}^2$ is the residual risk of the ith asset (the variance of the ith asset's returns left unexplained by the factors).

The direct procedure for calculating active factor risk is as follows. A portfolio's active factor exposure to a given factor j, b_j^a, is found by weighting each asset's sensitivity to factor j by its active weight and summing the terms:

$$b_j^a = \sum_{i=1}^{n} w_i^a b_{ji}.$$

Then active factor risk equals

$$\sum_{i=1}^{K}\sum_{j=1}^{K} b_i^a b_j^a \operatorname{cov}\left(F_i, F_j\right).$$

EXAMPLE 9

A Comparison of Active Risk

Richard Gray is comparing the risk of four US equity managers who share the same benchmark. He uses a fundamental factor model, the BARRA US-E4 model, which incorporates 12 style factors and a set of 60 industry factors. The style factors measure various fundamental aspects of companies and their shares, such as size, liquidity, leverage, and dividend yield. In the model, companies have non-zero exposures to all industries in which the company operates. Exhibit 10 presents Gray's analysis of the active risk squared of the four managers, based on Equation 12 (note that there is a covariance term in active factor risk, reflecting the correlation of industry membership and the risk indexes, which we assume is negligible in this example). In Exhibit 10, the column labeled "Industry" gives the portfolio's active factor risk associated with the industry exposures of its holdings; the "Style Factor" column gives the portfolio's active factor risk associated with the exposures of its holdings to the 12 style factors.

Exhibit 10: Active Risk Squared Decomposition

	Active Factor				
Portfolio	Industry	Style Factor	Total Factor	Active Specific	Active Risk Squared
A	12.25	17.15	29.40	19.60	49
B	1.25	13.75	15.00	10.00	25
C	1.25	17.50	18.75	6.25	25
D	0.03	0.47	0.50	0.50	1

Note: Entries are in % squared.

Using the information in Exhibit 10, address the following:

1. Contrast the active risk decomposition of Portfolios A and B.

Solution:

Exhibit 11 restates the information in Exhibit 10 to show the proportional contributions of the various sources of active risk. (e.g., Portfolio A's active risk related to industry exposures is 25% of active risk squared, calculated as 12.25/49 = 0.25, or 25%).

The last column of Exhibit 11 now shows the square root of active risk squared—that is, active risk or tracking error.

Exhibit 11: Active Risk Decomposition (restated)

Portfolio	Active Factor (% of total active)			Active Specific (% of total active)	Active Risk
	Industry	Style Factor	Total Factor		
A	25%	35%	60%	40%	7%
B	5%	55%	60%	40%	5%
C	5%	70%	75%	25%	5%
D	3%	47%	50%	50%	1%

Portfolio A has assumed a higher level of active risk than B (7% versus 5%). Portfolios A and B assumed the same proportions of active factor and active specific risk, but a sharp contrast exists between the two in the types of active factor risk exposure. Portfolio A assumed substantial active industry risk, whereas Portfolio B was approximately industry neutral relative to the benchmark. By contrast, Portfolio B had higher active bets on the style factors representing company and share characteristics.

2. Contrast the active risk decomposition of Portfolios B and C.

Solution:

Portfolios B and C were similar in their absolute amounts of active risk. Furthermore, both Portfolios B and C were both approximately industry neutral relative to the benchmark. Portfolio C assumed more active factor risk related to the style factors, but B assumed more active specific risk. It is also possible to infer from the greater level of B's active specific risk that B is somewhat less diversified than C.

3. Characterize the investment approach of Portfolio D.

Solution:

Portfolio D appears to be a passively managed portfolio, judging by its negligible level of active risk. Referring to Exhibit 11, Portfolio D's active factor risk of 0.50, equal to 0.707% expressed as a standard deviation, indicates that the portfolio's risk exposures very closely match the benchmark.

The discussion of performance attribution and risk analysis has used examples related to common stock portfolios. Multifactor models have also been effectively used in similar roles for portfolios of bonds and other asset classes. For example, such factors as duration and spread can be used to decompose the risk and return of a fixed-income manager.

FACTOR MODELS IN PORTFOLIO CONSTRUCTION

8

- ☐ describe uses of multifactor models and interpret the output of analyses based on multifactor models
- ☐ describe the potential benefits for investors in considering multiple risk dimensions when modeling asset returns

Equally as important to the use of multifactor models in analyzing a portfolio's active returns and active risk is the use of such multifactor models in portfolio construction. At this stage of the portfolio management process, multifactor models permit the portfolio manager to make focused bets or to control portfolio risk relative to the benchmark's risk. This greater level of detail in modeling risk that multifactor models afford is useful in both passive and active management.

- *Passive management.* In managing a fund that seeks to track an index with many component securities, portfolio managers may need to select a sample of securities from the index. Analysts can use multifactor models to replicate an index fund's factor exposures, mirroring those of the index tracked.
- *Active management.* Many quantitative investment managers rely on multifactor models in predicting alpha (excess risk-adjusted returns) or relative return (the return on one asset or asset class relative to that of another) as part of a variety of active investment strategies. In constructing portfolios, analysts use multifactor models to establish desired risk profiles.
- *Rules-based active management (alternative indexes).* These strategies routinely tilt toward such factors as size, value, quality, or momentum when constructing portfolios. As such, alternative index approaches aim to capture some systematic exposure traditionally attributed to manager skill, or "alpha," in a transparent, mechanical, rules-based manner at low cost. Alternative index strategies rely heavily on factor models to introduce intentional factor and style biases versus capitalization-weighted indexes.

In the following, we explore some of these uses in more detail. As indicated, an important use of multifactor models is to establish a specific desired risk profile for a portfolio. In the simplest instance, the portfolio manager may want to create a portfolio with sensitivity to a single factor. This particular (pure) factor portfolio would have a sensitivity of 1 for that factor and a sensitivity (or weight) of 0 for all other factors. It is thus a portfolio with exposure to only one risk factor and exactly represents the risk of that factor. As a pure bet on a source of risk, factor portfolios are of interest to a portfolio manager who wants to hedge that risk (offset it) or speculate on it. This simple case can be expanded to multiple factors where a factor replication portfolio can be built based either on an existing target portfolio or on a set of desired exposures. Example 10 illustrates the use of factor portfolios.

EXAMPLE 10

Factor Portfolios

Analyst Wanda Smithfield has constructed six portfolios for possible use by portfolio managers in her firm. The portfolios are labeled A, B, C, D, E, and F in Exhibit 12. Smithfield adapts a macroeconomic factor model based on research presented in Burmeister, Roll, and Ross (1994). The model includes five factors:

- Confidence risk, based on the yield spread between corporate bonds and government bonds. A positive surprise in the spread suggests that investors are willing to accept a smaller reward for bearing default risk and so that confidence is high.
- Time horizon risk, based on the yield spread between 20-year government bonds and 30-day Treasury bills. A positive surprise indicates increased investor willingness to invest for the long term.
- Inflation risk, measured by the unanticipated change in the inflation rate.
- Business cycle risk, measured by the unexpected change in the level of real business activity.
- Market timing risk, measured as the portion of the return on a broad-based equity index that is unexplained by the first four risk factors.

Exhibit 12: Factor Portfolios

	Portfolios					
Risk Factor	**A**	**B**	**C**	**D**	**E**	**F**
Confidence risk	0.50	0.00	1.00	0.00	0.00	0.80
Time horizon risk	1.92	0.00	1.00	1.00	1.00	1.00
Inflation risk	0.00	0.00	1.00	0.00	0.00	−1.05
Business cycle risk	1.00	1.00	0.00	0.00	1.00	0.30
Market timing risk	0.90	0.00	1.00	0.00	0.00	0.75

Note: Entries are factor sensitivities.

1. A portfolio manager wants to place a bet that real business activity will increase.

   Determine and justify the portfolio among the six given that would be most useful to the manager.

Solution:

Portfolio B is the most appropriate choice. Portfolio B is the factor portfolio for business cycle risk because it has a sensitivity of 1 to business cycle risk and a sensitivity of 0 to all other risk factors. Portfolio B is thus efficient for placing a pure bet on an increase in real business activity.

2. Would the manager take a long or short position in the portfolio chosen in Part A?

Solution:

The manager would take a long position in Portfolio B to place a bet on an increase in real business activity.

3. A portfolio manager wants to hedge an existing positive (long) exposure to time horizon risk.

Determine and justify the portfolio among the six given that would be most useful to the manager.

Solution:

Portfolio D is the appropriate choice. Portfolio D is the factor portfolio for time horizon risk because it has a sensitivity of 1 to time horizon risk and a sensitivity of 0 to all other risk factors. Portfolio D is thus efficient for hedging an existing positive exposure to time horizon risk.

4. What type of position would the manager take in the portfolio chosen in Part A?

Solution:

The manager would take a short position in Portfolio D to hedge the positive exposure to time horizon risk.

CONSTRUCTING MULTIFACTOR PORTFOLIOS

In practice, most stock selection models use some common multifactor structure. Here, we describe constructing two types of multifactor portfolios—a benchmark portfolio and a risk parity portfolio—that target desired risk exposures to eight fundamental factors. The benchmark portfolio equally weights the pure factors, whereas the risk parity portfolio weights the pure factors based on equal risk contribution. We focus on the benchmark and risk parity portfolios because their factor weighting schemes are clear and objective.

Setting the Scene: Pure Factor Portfolios

For demonstration purposes, we use fundamental factor models and choose common company- and company share–related factors from each main investment style (i.e., value, growth, price momentum, analyst sentiment, and quality):

1. *Defensive value*: Trailing earnings yield—companies with high earnings yield are preferred.
2. *Cyclical value*: Book-to-market ratio—companies with high book-to-market ratios (i.e., cheap stock valuations) are bought.
3. *Growth*: Consensus FY1/FY0 EPS growth—companies with high expected earnings growth are preferred.
4. *Price momentum*: 12M total return excluding the most recent month—companies with positive price momentum are preferred.
5. *Analyst sentiment*: 3M EPS revision—companies with positive earnings revisions are bought.

6. *Profitability*: Return on equity (ROE)—companies with high ROEs are bought.
7. *Leverage*: Debt/equity ratio—companies with low financial leverage are preferred.
8. *Earnings quality*: Non-cash earnings—companies with low accruals are bought. Research suggests that net income with low levels of non-cash items (i.e., accruals) is less likely to be manipulated.

The stock universe for this demonstration consists of the Russell 3000 Index (US), the S&P/TSX Composite Index (Canada), the MSCI China A Index (China), and the S&P Global Broad Market Index (all other countries). A pure factor portfolio is formed for each of the eight factors by buying the top 20% of stocks and shorting the bottom 20% of stocks ranked by the factor. Stocks held long and short are equally weighted, and the eight factor portfolios are each rebalanced monthly. Note that this demonstration does not account for transaction costs or other portfolio constraints. Other methods for forming pure factor portfolios include ranking stocks by Pearson IC (correlation between prior period factor scores and current period stock returns) or by Spearman Rank IC (correlation between prior period ranked factor scores and current period ranked stock returns), as well as ranking by other univariate regression methods. However, for simplicity, we follow the long–short portfolio approach.

A straightforward way to combine these pure factor portfolios into a multifactor portfolio is equal weighting. We call the equally weighted multifactor portfolio the "benchmark (BM) portfolio." The experience in practice is that portfolios constructed using this simple weighting scheme typically perform at least as well as those using more sophisticated optimization techniques.

Risk parity is a common alternative portfolio construction technique used in the asset allocation space. Risk parity accounts for the volatility of each factor and the correlations of returns among all factors to be combined into the multifactor portfolio. The objective is for each factor to contribute equally to the overall (or targeted) risk of the portfolio. Thus, a risk parity (RP) multifactor portfolio can be created by equally weighting the risk contribution of each of the eight pure factors mentioned.

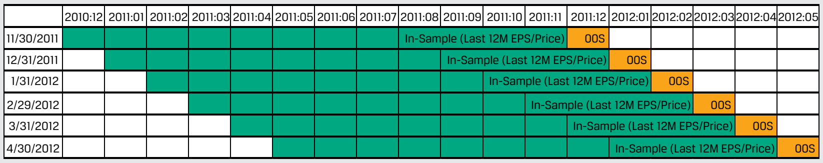
Constructing and Backtesting Benchmark and Risk Parity Multifactor Portfolios

To create a successful multifactor portfolio strategy, the investment manager needs to perform backtesting to assess factor performance and effectiveness. In a typical backtest, a manager first forms her investment hypothesis, determines her investment rules and processes, collects the required data, and creates the portfolio, and then she periodically rebalances and evaluates the portfolio.

In the rolling window backtesting methodology, analysts use a rolling window framework, fit factors based on the rolling window, rebalance the portfolio periodically, and then track performance. Thus, backtesting is a proxy for actual investing. As new information arrives, investment managers readjust their models and rebalance their stock positions, typically monthly. Thus, they repeat the same in-sample training/out-of-sample testing process. If the investment strategy's performance in out-of-sample periods is desirable and the strategy makes intuitive sense, then it is deemed successful.

The following exhibit illustrates rolling window backtesting of the defensive value factor from November 2011 to April 2012. On 30 November 2011, we compute each stock's trailing 12-month earnings yield, then buy the 20% of stocks with the highest earnings yield and short the bottom quintile of stocks, and assess performance using returns in the next month, December 2011, the out-of-sample (OOS) period. The process is repeated on 31 December 2011, and so on, and finally, we compute the average monthly return, volatility, Sharpe ratio, and drawdown from the test results of the six OOS periods.

An Example of Rolling Window Backtesting of the Defensive Value Factor



Source: Wolfe Research Luo's QES.

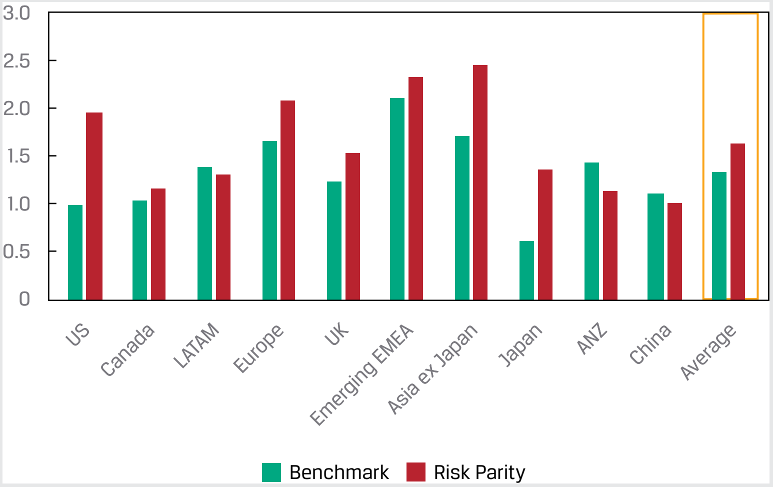
Constructing and backtesting multifactor portfolios is similar to the method just described, except that the rolling window procedure is implemented twice. First, we form the eight pure factor portfolios for each month from 1988 until May 2019 by implementing the rolling window procedure. Then, we combine the underlying factor portfolios into the multifactor portfolios using the two approaches—equally weighting all factors (i.e., benchmark, or BM, allocation) and equally risk weighting all factors (i.e., risk parity, or RP, allocation).

Importantly, the process for creating the multifactor portfolios requires a second implementation of the rolling window procedure to avoid look-ahead bias; note this second rolling window covers the same time span as the first one (i.e., 1988 until May 2019). At each month-end, the previous five years of monthly data are used to estimate the variance–covariance matrix for the eight factor portfolios. Once the covariance matrix is estimated, we optimize and compute the weights for each of the eight pure factor portfolios and then form the RP portfolio. Finally, we compute the returns of the two multifactor portfolios (BM and RP) during this out-of-sample period using the weights at the end of the previous month and the returns of the eight underlying factor portfolios for the current month. This process is repeated every month over the entire horizon of 1988 until May 2019.

We created and backtested the multifactor portfolios using both the equal weighting (BM) scheme and risk parity (RP) scheme for each of 10 markets, including the United States. Both multifactor portfolios are rebalanced monthly to maintain equal factor weights or equal factor risk contributions. As noted previously, the key input to the RP allocation is the monthly variance–covariance matrix for the eight underlying factor portfolios derived from the rolling (five-year) window procedure. To be clear, each of the eight factor portfolios is a long–short portfolio. However, our factor allocation strategies to form the BM and RP multifactor portfolios are long only, meaning the weights allocated to each of the eight factor portfolios are restricted to be non-negative. Therefore, factor weights for the BM and RP portfolios are positive and add to 100%.

In the United States over the period 1993–2019, the weights of the eight factor portfolios in the RP allocation are relatively stable. Interestingly, book-to-market and earnings quality factor portfolios receive the largest allocations, whereas ROE and price momentum factor portfolios have the lowest weights. The RP multifactor portfolio provides a lower cumulative return than does the BM multifactor portfolio; however, the RP portfolio's volatility is substantially lower than that of the BM portfolio. Consequently, in the United States, the RP portfolio's Sharpe ratio is nearly double that of the BM portfolio, as shown in the following exhibit. Outperformance of the RP portfolio in terms of Sharpe ratio is also apparent across most markets examined.

Average Sharpe Ratios for Multifactor Portfolios: Equally Weighted vs. Risk Parity Weighted (1993–2019)



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES.

The following case was written by Yin Luo, CPA, PStat, CFA, and Sheng Wang, both of Wolfe Research LLC (USA).

9 FACTOR MODELS IN STRATEGIC PORTFOLIO DECISIONS

- ☐ describe uses of multifactor models and interpret the output of analyses based on multifactor models
- ☐ describe the potential benefits for investors in considering multiple risk dimensions when modeling asset returns

Multifactor models can help investors recognize considerations that are relevant in making various strategic decisions. For example, given a sound model of the systematic risk factors that affect assets' mean returns, the investor can ask, relative to other investors,

- What types of risk do I have a comparative advantage in bearing?
- What types of risk am I at a comparative disadvantage in bearing?

For example, university endowments, because they typically have very long investment horizons, may have a comparative advantage in bearing business cycle risk of traded equities or the liquidity risk associated with many private equity investments. They may tilt their strategic asset allocation or investments within an asset class to capture the associated risk premiums for risks that do not much affect them. However, such investors may be at a comparative disadvantage in bearing inflation risk to the extent that the activities they support have historically been subject to cost increases running above the average rate of inflation.

This is a richer framework than that afforded by the CAPM, according to which all investors optimally should invest in two funds: the market portfolio and a risk-free asset. Practically speaking, a CAPM-oriented investor might hold a money market fund and a portfolio of capitalization-weighted broad market indexes across many asset classes, varying the weights in these two in accordance with risk tolerance. These types of considerations are also relevant to individual investors. An individual investor who depends on income from salary or self-employment is sensitive to business cycle risk, in particular to the effects of recessions. If this investor compared two stocks with the same CAPM beta, given his concern about recessions, he might be very sensitive to receiving an adequate premium for investing in procyclical assets. In contrast, an investor with independent wealth and no job-loss concerns would have a comparative advantage in bearing business cycle risk; his optimal risky asset portfolio might be quite different from that of the investor with job-loss concerns in tilting toward greater-than-average exposure to the business cycle factor, all else being equal. Investors should be aware of which priced risks they face and analyze the extent of their exposure.

A multifactor approach can help investors achieve better-diversified and possibly more-efficient portfolios. For example, the characteristics of a portfolio can be better explained by a combination of SMB, HML, and WML factors in addition to the market factor than by using the market factor alone.

Thus, compared with single-factor models, multifactor models offer a richer context for investors to search for ways to improve portfolio selection.

SUMMARY

In our coverage of multifactor models, we have presented concepts, models, and tools that are key ingredients to quantitative portfolio management and are used to both construct portfolios and to attribute sources of risk and return.

- Multifactor models permit a nuanced view of risk that is more granular than the single-factor approach allows.
- Multifactor models describe the return on an asset in terms of the risk of the asset with respect to a set of factors. Such models generally include systematic factors, which explain the average returns of a large number of risky assets. Such factors represent priced risk—risk for which investors require an additional return for bearing.
- The arbitrage pricing theory (APT) describes the expected return on an asset (or portfolio) as a linear function of the risk of the asset with respect to a set of factors. Like the CAPM, the APT describes a financial market equilibrium; however, the APT makes less strong assumptions.
- The major assumptions of the APT are as follows:
    - Asset returns are described by a factor model.
    - With many assets to choose from, asset-specific risk can be eliminated.
    - Assets are priced such that there are no arbitrage opportunities.
- Multifactor models are broadly categorized according to the type of factor used:
    - Macroeconomic factor models
    - Fundamental factor models
    - Statistical factor models

- In *macroeconomic* factor models, the factors are surprises in macroeconomic variables that significantly explain asset class (equity in our examples) returns. Surprise is defined as actual minus forecasted value and has an expected value of zero. The factors can be understood as affecting either the expected future cash flows of companies or the interest rate used to discount these cash flows back to the present and are meant to be uncorrelated.
- In *fundamental* factor models, the factors are attributes of stocks or companies that are important in explaining cross-sectional differences in stock prices. Among the fundamental factors are book-value-to-price ratio, market capitalization, price-to-earnings ratio, and financial leverage.
- In contrast to macroeconomic factor models, in fundamental models the factors are calculated as returns rather than surprises. In fundamental factor models, we generally specify the factor sensitivities (attributes) first and then estimate the factor returns through regressions. In macroeconomic factor models, however, we first develop the factor (surprise) series and then estimate the factor sensitivities through regressions. The factors of most fundamental factor models may be classified as company fundamental factors, company share-related factors, or macroeconomic factors.
- In *statistical* factor models, statistical methods are applied to a set of historical returns to determine portfolios that explain historical returns in one of two senses. In factor analysis models, the factors are the portfolios that best explain (reproduce) historical return covariances. In principal-components models, the factors are portfolios that best explain (reproduce) the historical return variances.
- Multifactor models have applications to return attribution, risk attribution, portfolio construction, and strategic investment decisions.
- A factor portfolio is a portfolio with unit sensitivity to a factor and zero sensitivity to other factors.
- Active return is the return in excess of the return on the benchmark.
- Active risk is the standard deviation of active returns. Active risk is also called tracking error or tracking risk. Active risk squared can be decomposed as the sum of active factor risk and active specific risk.
- The information ratio (IR) is mean active return divided by active risk (tracking error). The IR measures the increment in mean active return per unit of active risk.
- Factor models have uses in constructing portfolios that track market indexes and in alternative index construction.
- Traditionally, the CAPM approach would allocate assets between the risk-free asset and a broadly diversified index fund. Considering multiple sources of systematic risk may allow investors to improve on that result by tilting away from the market portfolio. Generally, investors would gain from accepting above average (below average) exposures to risks that they have a comparative advantage (comparative disadvantage) in bearing.

REFERENCES

Bodie, Zvi, Alex Kane, and Alan J. Marcus. 2017. *Investments*. 11th ed.New York: McGraw-Hill Education.

Burmeister, Edwin, Richard Roll, and Stephen A. Ross. 1994. "A *Practitioner's* Guide to Arbitrage Pricing Theory." In *A Practitioner's Guide to Factor Models*. Charlottesville, VA: Research Foundation of the Institute of Chartered Financial Analysts.

Carhart, Mark M. 1997. "On Persistence in Mutual Fund Performance." *Journal of Finance* 52 (1): 57–82. 10.1111/j.1540-6261.1997.tb03808.x

Connor, Gregory. 1995. "The Three Types of Factor Models: A Comparison of Their Explanatory Power." *Financial Analysts Journal* 51 (3): 42–46. 10.2469/faj.v51.n3.1904

Dopfel, Frederick E. 2004. "Fixed-Income Style Analysis and Optimal Manager Structure." *Journal of Fixed Income* 14 (2): 32–43. 10.3905/jfi.2004.439835

Fabozzi, Frank J. 2008. *Handbook of Finance, Financial Markets and Instruments*. Wiley.

Fama, Eugene F. and Kenneth R. French. 1992. "The Cross-Section of Expected Stock Returns." *Journal of Finance* 47 (2): 427–65. 10.1111/j.1540-6261.1992.tb04398.x

Grinold, Richard and Ronald N. Kahn. 1994. "Multi-Factor Models for Portfolio Risk." In *A Practitioner's Guide to Factor Models*. Charlottesville, VA: Research Foundation of the Institute of Chartered Financial Analysts.

Podkaminer, Eugene L. 2017. "Smart Beta Is the Gateway Drug to Risk Factor Investing." *Journal of Portfolio Management* 43 (5): 130–34. 10.3905/jpm.2017.43.5.130

Ross, S. A. 1976. "The Arbitrage Theory of Capital Asset Pricing." *Journal of Economic Theory* 13 (3): 341–60. 10.1016/0022-0531(76)90046-6

PRACTICE PROBLEMS

1. Compare the assumptions of the arbitrage pricing theory (APT) with those of the capital asset pricing model (CAPM).

2. Assume that the following one-factor model describes the expected return for portfolios:

   $E(R_p) = 0.10 + 0.12\beta_{p,1}$

   Also assume that all investors agree on the expected returns and factor sensitivity of the three highly diversified Portfolios A, B, and C given in the following table:

Portfolio	Expected Return	Factor Sensitivity
A	0.196	0.80
B	0.156	1.00
C	0.244	1.20

   Assuming the one-factor model is correct and based on the data provided for Portfolios A, B, and C, determine if an arbitrage opportunity exists and explain how it might be exploited.

The following information relates to questions 3-8

Carlos Altuve is a manager-of-managers at an investment company that uses quantitative models extensively. Altuve seeks to construct a multi-manager portfolio using some of the funds managed by portfolio managers within the firm. Maya Zapata is assisting him.

Altuve uses arbitrage pricing theory (APT) as a basis for evaluating strategies and managing risks. From his earlier analysis, Zapata knows that Funds A and B in Exhibit 1 are well diversified. He has not previously worked with Fund C and is puzzled by the data because it is inconsistent with APT. He asks Zapata gather additional information on Fund C's holdings and to determine if an arbitrage opportunity exists among these three investment alternatives. Her analysis, using the data in Exhibit 1, confirms that an arbitrage opportunity does exist.

Exhibit 1: Expected Returns and Factor Sensitivities (One-Factor Model)

Fund	Expected Return	Factor Sensitivity
A	0.02	0.5
B	0.04	1.5
C	0.03	0.9

Using a two-factor model, Zapata now estimates the three funds' sensitivity to inflation and GDP growth. That information is presented in Exhibit 2. Zapata assumes a zero value for the error terms when working with the selected two-factor model.

Exhibit 2: Expected Returns and Factor Sensitivities (Two-Factor Model)

Fund	Expected Return	Factor Sensitivity	
		Inflation	GDP Growth
A	0.02	0.5	1.0
B	0.04	1.6	0.0
C	0.03	1.0	1.1

Altuve asks Zapata to calculate the return for Portfolio AC, composed of a 60% allocation to Fund A and 40% allocation to Fund C, using the surprises in inflation and GDP growth in Exhibit 3.

Exhibit 3: Selected Data on Factors

Factor	Research Staff Forecast	Actual Value
Inflation	2.0%	2.2%
GDP Growth	1.5%	1.0%

Finally, Altuve asks Zapata about the return sensitivities of Portfolios A, B, and C given the information provided in Exhibit 3.

3. Which of the following is *not* a key assumption of APT, which is used by Altuve to evaluate strategies and manage risks?
   - **A.** A factor model describes asset returns.
   - **B.** Asset-specific risk can be eliminated through diversification.
   - **C.** Arbitrage opportunities exist among well-diversified portfolios.

4. The arbitrage opportunity identified by Zapata can be exploited with:
   - **A.** Strategy 1: Buy $50,000 Fund A and $50,000 Fund B; sell short $100,000 Fund C.
   - **B.** Strategy 2: Buy $60,000 Fund A and $40,000 Fund B; sell short $100,000 Fund C.
   - **C.** Strategy 3: Sell short $60,000 of Fund A and $40,000 of Fund B; buy $100,000 Fund C.

5. The two-factor model Zapata uses is a:
   - **A.** statistical factor model.
   - **B.** fundamental factor model.
   - **C.** macroeconomic factor model.

6. Based on the data in Exhibits 2 and 3, the return for Portfolio AC, given the sur-

prises in inflation and GDP growth, is *closest* to:

A. 2.02%.

B. 2.40%.

C. 4.98%.

7. The surprise in which of the following had the greatest effect on fund returns?

A. Inflation on Fund B

B. GDP growth on Fund A

C. GDP growth on Fund C

8. Based on the data in Exhibit 2, which fund is most sensitive to the combined surprises in inflation and GDP growth in Exhibit 3?

A. Fund A

B. Fund B

C. Fund C

The following information relates to questions 9-14

Hui Cheung, a portfolio manager, asks her assistant, Ronald Lam, to review the macroeconomic factor model currently in use and to consider a fundamental factor model as an alternative.

The current macroeconomic factor model has four factors:

$$R_i = a_i + b_{i1}F_{\text{GDP}} + b_{i2}F_{\text{CAP}} + b_{i3}F_{\text{CON}} + b_{i4}F_{\text{UNEM}} + \varepsilon_i,$$

Where F_{GDP}, F_{CAP}, F_{CON}, and F_{UNEM} represent unanticipated changes in four factors: gross domestic product, manufacturing capacity utilization, consumer spending, and the rate of unemployment, respectively. Lam assumes the error term is equal to zero when using this model.

Lam estimates the current model using historical monthly returns for three portfolios for the most recent five years. The inputs used in and estimates derived from the macroeconomic factor model are presented in Exhibit 1. The US Treasury bond rate of 2.5% is used as a proxy for the risk-free rate of interest.

Exhibit 1: Inputs for and Estimates from the Current Macroeconomic Model

	Factor Sensitivities and Intercept Coefficients				
Factor	**Portfolio 1**	**Portfolio 2**	**Portfolio 3**	**Benchmark**	**Factor Surprise (%)**
Intercept (%)	2.58	3.20	4.33		
F_{GDP}	0.75	1.00	0.24	0.50	0.8
F_{CAP}	−0.23	0.00	−1.45	−1.00	0.5

Factor	Factor Sensitivities and Intercept Coefficients: Portfolio 1	Portfolio 2	Portfolio 3	Benchmark	Factor Surprise (%)
F_{CON}	1.23	0.00	0.50	1.10	2.5
F_{UNEM}	−0.14	0.00	−0.05	−0.10	1.0
Annual Returns, Most Recent Year					
Return (%)	6.00	4.00	5.00	4.50	

Lam uses the macroeconomic model to calculate the tracking error and the mean active return for each portfolio. He presents these statistics in Exhibit 2.

Exhibit 2: Macroeconomic Factor Model Tracking Error and Mean Active Return

Portfolio	Tracking Error	Mean Active Return
Portfolio 1	1.50%	1.50%
Portfolio 2	1.30%	−0.50%
Portfolio 3	1.00%	0.50%

Lam considers a fundamental factor model with four factors:

$$R_i = a_j + b_{j1}F_{LIQ} + b_{j2}F_{LEV} + b_{j3}F_{EGR} + b_{j4}F_{VAR} + \varepsilon_j,$$

where F_{LIQ}, F_{LEV}, F_{EGR}, and F_{VAR} represent liquidity, financial leverage, earnings growth, and the variability of revenues, respectively.

Lam and Cheung discuss similarities and differences between macroeconomic factor models and fundamental factor models, and Lam offers a comparison of those models to statistical factor models. Lam makes the following statements.

Statement 1 The factors in fundamental factor models are based on attributes of stocks or companies, whereas the factors in macroeconomic factor models are based on surprises in economic variables.

Statement 2 The factor sensitivities are generally determined first in fundamental factor models, whereas the factor sensitivities are estimated last in macroeconomic factor models.

Lam also tells Cheung:

An advantage of statistical factor models is that they make minimal assumptions, and therefore, statistical factor model estimation lends itself to easier interpretation than macroeconomic and fundamental factor models.

Lam tells Cheung that multifactor models can be useful in active portfolio management, but not in passive management. Cheung disagrees; she tells Lam that multifactor models can be useful in both active and passive management.

9. Based on the information in Exhibit 1, the expected return for Portfolio 1 is *closest* to:

A. 2.58%.

B. 3.42%.

C. 6.00%.

10. Based on Exhibit 1, the active risk for Portfolio 2 is explained by surprises in:

A. GDP.

B. consumer spending.

C. all four model factors.

11. Based on Exhibit 2, which portfolio has the best information ratio?

A. Portfolio 1

B. Portfolio 2

C. Portfolio 3

12. Which of Lam's statements regarding macroeconomic factor models and fundamental factor models is correct?

A. Only Statement 1

B. Only Statement 2

C. Both Statements 1 and 2

13. Is Lam's comment regarding statistical factor models correct?

A. Yes

B. No, because he is incorrect with respect to interpretation of the models' results

C. No, because he is incorrect with respect to the models' assumptions

14. Whose statement regarding the use of multifactor models in active and passive portfolio management is correct?

A. Lam only

B. Cheung only

C. Both Lam and Cheung

---

15. Last year the return on Harry Company stock was 5 percent. The portion of the return on the stock not explained by a two-factor macroeconomic factor model was 3 percent. Using the data given below, calculate Harry Company stock's expected return.

Macroeconomic Factor Model for Harry Company Stock

Variable	Actual Value (%)	Expected Value (%)	Stock's Factor Sensitivity
Change in interest rate	2.0	0.0	–1.5
Growth in GDP	1.0	4.0	2.0

16. Which type of factor model is most directly applicable to an analysis of the style orientation (for example, growth vs. value) of an active equity investment manager? Justify your answer.

17. Suppose an active equity manager has earned an active return of 110 basis points, of which 80 basis points is the result of security selection ability. Explain the likely source of the remaining 30 basis points of active return.

The following information relates to questions 18-19

Address the following questions about the information ratio.

18. What is the information ratio of an index fund that effectively meets its investment objective?

19. What are the two types of risk an active investment manager can assume in seeking to increase his information ratio?

---

20. A wealthy investor has no other source of income beyond her investments and that income is expected to reliably meet all her needs. Her investment advisor recommends that she tilt her portfolio to cyclical stocks and high-yield bonds. Explain the advisor's advice in terms of comparative advantage in bearing risk.

SOLUTIONS

1. APT and the CAPM are both models that describe what the expected return on a risky asset should be in equilibrium given its risk. The CAPM is based on a set of assumptions including the assumption that investors' portfolio decisions can be made considering just returns' means, variances, and correlations. The APT makes three assumptions:

   1. A factor model describes asset returns.
   2. There are many assets, so investors can form well-diversified portfolios that eliminate asset-specific risk.
   3. No arbitrage opportunities exist among well-diversified portfolios.

2. According to the one-factor model for expected returns, the portfolio should have these expected returns if they are correctly priced in terms of their risk:

   Portfolio A: $E(R_A) = 0.10 + 0.12\beta_{A,1} = 0.10 + (0.12)(0.80) = 0.10 + 0.10 = 0.20$

   Portfolio B: $E(R_B) = 0.10 + 0.12\beta_{B,1} = 0.10 + (0.12)(1.00) = 0.10 + 0.12 = 0.22$

   Portfolio C: $E(R_C) = 0.10 + 0.12\beta_{C,1} = 0.10 + (0.12)(1.20) = 0.10 + 0.14 = 0.24$

   In the table below, the column for expected return shows that Portfolios A and C are correctly priced but Portfolio B offers too little expected return for its risk, 0.15 or 15%. By shorting Portfolio B (selling an overvalued portfolio) and using the proceeds to buy a portfolio 50% invested in A and 50% invested in C with a sensitivity of 1 that matches the sensitivity of B, for each monetary unit shorted (say each euro), an arbitrage profit of €0.22 – €0.15 = €0.07 is earned.

Portfolio	Expected Return	Factor Sensitivity
A	0.196	0.80
B	0.156	1.00
C	0.244	1.20
0.5A + 0.5C	0.22	1.00

3. C is correct. Arbitrage pricing theory (APT) is a framework that explains the expected return of a portfolio in equilibrium as a linear function of the risk of the portfolio with respect to a set of factors capturing systematic risk. A key assumption of APT is that, in equilibrium, there are no arbitrage opportunities.

4. C is correct. The expected return and factor sensitivities of a portfolio with a 60% weight in Fund A and a 40% weight in Fund B are calculated as weighted averages of the expected returns and factor sensitivities of Funds A and B:

$$\text{Expected return of Portfolio 60/40} = (0.60)(0.02) + (0.40)(0.04) = 0.028\text{, or }2.8\%$$

$$\text{Factor sensitivity of Portfolio 60/40} = (0.60)(0.5) + (0.40)(1.5) = 0.9$$

Fund	Expected Return	Factor Sensitivity
A	0.02	0.5
B	0.04	1.5
C	0.03	0.9

Portfolio 60/40		
60%A + 40%B	0.028	0.900

Portfolio 50/50		
50%A + 50%B	0.030	1.000

The factor sensitivity of Portfolio 60/40 is identical to that of Fund C; therefore, this strategy results in no factor risk relative to Portfolio C. However, Fund C's expected return of 3.0% is higher than Portfolio 60/40's expected return of 2.8%. This difference supports Strategy 3: buying Fund C and selling short Portfolio 60/40 to exploit the arbitrage opportunity.

5. C is correct. In a macroeconomic factor model, the factors are surprises in macroeconomic variables, such as inflation risk and GDP growth, that significantly explain returns.

6. A is correct. The macroeconomic two-factor model takes the following form:

$$R_i = a_i + b_{i1}F_{\text{INF}} + b_{i2}F_{\text{GDP}} + \varepsilon_i,$$

where F_{INF} and F_{GDP} represent surprises in inflation and surprises in GDP growth, respectively, and a_i represents the expected return to asset i. Using this model and the data in Exhibit 2, the returns for Fund A and Fund C are represented by the following:

$$R_A = 0.02 + 0.5F_{\text{INF}} + 1.0F_{\text{GDP}} + \varepsilon_A$$

$$R_C = 0.03 + 1.0F_{\text{INF}} + 1.1F_{\text{GDP}} + \varepsilon_C$$

Surprise in a macroeconomic model is defined as actual factor minus predicted factor. The surprise in inflation is 0.2% (= 2.2% – 2.0%). The surprise in GDP growth is –0.5% (= 1.0% – 1.5%). The return for Portfolio AC, composed of a 60% allocation to Fund A and 40% allocation to Fund C, is calculated as the following:

$$R_{AC} = (0.6)(0.02) + (0.4)(0.03) + [(0.6)(0.5) + (0.4)(1.0)](0.002) + [(0.6)(1.0) + (0.4)(1.1)](-0.005) + 0.6(0) + 0.4(0)$$

$$= 2.02\%$$

7. C is correct. Surprise in a macroeconomic model is defined as actual factor minus predicted factor. For inflation, the surprise factor is 2.2% – 2.0% = 0.2%; for GDP growth, the surprise factor is 1.0% – 1.5% = –0.5%. The effect on returns is the product of the surprise and the factor sensitivity.

	Change in Portfolio Return due to Surprise in	
Fund	**Inflation**	**GDP Growth**
A	0.5 × 0.2% = 0.10%	1.0 × −0.5% = −0.50%
B	1.6 × 0.2% = 0.32%	0.0 × −0.5% = 0.00%
C	1.0 × 0.2% = 0.20%	1.1 × −0.5% = −0.55%

The effect of the GDP growth surprise on Fund C was the largest single-factor effect on Fund returns (−0.55%).

8. A is correct. The effect of the surprises in inflation and GDP growth on the returns of the three funds is calculated as the following.

	Change in Portfolio Return Because of Surprise in	
Fund	**Inflation**	**GDP Growth**
A	0.5 × 0.2% = 0.10%	1.0 × −0.5% = −0.50%
B	1.6 × 0.2% = 0.32%	0.0 × −0.5% = 0.00%
C	1.0 × 0.2% = 0.20%	1.1 × −0.5% = −0.55%

The combined effects for the three funds are the following.

Fund A: 0.10% + (−0.50%) = −0.40%

Fund B: 0.32% + (0.00%) = 0.32%

Fund C: 0.20% + (−0.55%) = −0.35%

Therefore, Fund A is the most sensitive to the surprises in inflation and GDP growth in Exhibit 3.

9. A is correct. When using a macroeconomic factor model, the expected return is the intercept (when all model factors take on a value of zero). The intercept coefficient for Portfolio 1 in Exhibit 1 is 2.58.

10. C is correct. Active risk, also referred to as tracking risk or tracking error, is the sample standard deviation of the time series of active returns, where the active returns consist of the differences between the portfolio return and the benchmark return. Whereas GDP is the only portfolio non-zero sensitivity for Portfolio 2, the contribution to the portfolio's active return is the sum of the differences between the portfolio's and the benchmark's sensitivities multiplied by the factor return. Because all four of the factor sensitivities of Portfolio 2 are different from the factor sensitivities of the benchmark, all four factors contribute to the portfolio's active return and, therefore, to its active risk.

11. A is correct. Portfolio 1 has the highest information ratio, 1.0, and thus has the best mean active return per unit of active risk:

$$\begin{aligned} \text{IR} &= \frac{\overline{R}_P - \overline{R}_B}{s\left(R_P - R_B\right)} \\ &= \frac{1.50\%}{1.50\%} \\ &= 1.00 \end{aligned}$$

This information ratio exceeds that of Portfolio 2 (−0.38) or Portfolio 3 (0.50).

12. C is correct. In a macroeconomic factor model, the factors are surprises in macroeconomic variables that significantly explain returns. Factor sensitivities are

generally specified first in fundamental factor models, whereas factor sensitivities are estimated last in macroeconomic factor models.

13. B is correct. An advantage of statistical factor models is that they make minimal assumptions. However, the interpretation of statistical factors is generally more difficult than the interpretation of macroeconomic and fundamental factor models.

14. B is correct. Analysts can use multifactor models in passively managed portfolios to replicate an index fund's factor exposures.

15. In a macroeconomic factor model, the surprise in a factor equals actual value minus expected value. For the interest rate factor, the surprise was 2 percent; for the GDP factor, the surprise was −3 percent. The intercept represents expected return in this type of model. The portion of the stock's return not explained by the factor model is the model's error term.

$$5\% = \text{Expected return} - 1.5(\text{Interest rate surprise}) + 2(\text{GDP surprise}) + \text{Error term}$$

$$= \text{Expected return} - 1.5(2\%) + 2(-3\%) + 3\%$$

$$= \text{Expected return} - 6\%$$

Rearranging terms, the expected return for Harry Company stock equals 5% + 6% = 11%.

16. A fundamental factor model. Such models typically include many factors related to the company (e.g., earnings) and to valuation that are commonly used indicators of a growth orientation. A macroeconomic factor model may provide relevant information as well, but typically indirectly and in less detail.

17. This remainder of 30 basis points would be attributable to the return from factor tilts. A portfolio manager's active return is the sum of two components, factor tilts and security selection. Factor tilt is the product of the portfolio manager's higher or lower factor sensitivities relative to the benchmark's factor sensitivities and the factor returns. Security selection reflects the manager's ability to overweight securities that outperform or underweight securities that underperform.

18. An index fund that effectively meets its investment objective is expected to have an information ratio of zero, because its active return should be zero.

19. The active manager may assume active factor risk and active specific risk (security selection risk) in seeking a higher information ratio.

20. This wealthy investor has a comparative advantage in bearing business cycle risk compared with the average investor who depends on income from employment. Because the average investor is sensitive to the business cycle and in particular the risk of recession, we would expect there to be a risk premium to hold recession-sensitive securities. Cyclical stocks and high-yield bonds are both very sensitive to the risk of recessions. Because the welfare of the wealthy investor is not affected by recessions, she can tilt her portfolio to include cyclical stocks and high yield bonds to attempt to capture the associated risk premiums.

LEARNING MODULE

5

Measuring and Managing Market Risk

by Don M. Chance, PhD, CFA, and Michelle McCarthy Beck.

Don M. Chance, PhD, CFA, is at Louisiana State University (USA). Michelle McCarthy Beck is at TIAA (USA).

LEARNING OUTCOMES

Mastery	*The candidate should be able to:*
☐	explain the use of value at risk (VaR) in measuring portfolio risk
☐	compare the parametric (variance–covariance), historical simulation, and Monte Carlo simulation methods for estimating VaR
☐	estimate and interpret VaR under the parametric, historical simulation, and Monte Carlo simulation methods
☐	describe advantages and limitations of VaR
☐	describe extensions of VaR
☐	describe sensitivity risk measures and scenario risk measures and compare these measures to VaR
☐	demonstrate how equity, fixed-income, and options exposure measures may be used in measuring and managing market risk and volatility risk
☐	describe the use of sensitivity risk measures and scenario risk measures
☐	describe advantages and limitations of sensitivity risk measures and scenario risk measures
☐	explain constraints used in managing market risks, including risk budgeting, position limits, scenario limits, and stop-loss limits
☐	explain how risk measures may be used in capital allocation decisions
☐	describe risk measures used by banks, asset managers, pension funds, and insurers

1 INTRODUCTION

☐ explain the use of value at risk (VaR) in measuring portfolio risk

This reading is an introduction to the process of measuring and managing market risk. Market risk is the risk that arises from movements in stock prices, interest rates, exchange rates, and commodity prices. Market risk is distinguished from credit risk, which is the risk of loss from the failure of a counterparty to make a promised payment, and also from a number of other risks that organizations face, such as breakdowns in their operational procedures. In essence, market risk is the risk arising from changes in the markets to which an organization has exposure.

Risk management is the process of identifying and measuring risk and ensuring that the risks being taken are consistent with the desired risks. The process of managing market risk relies heavily on the use of models. A model is a simplified representation of a real world phenomenon. Financial models attempt to capture the important elements that determine prices and sensitivities in financial markets. In doing so, they provide critical information necessary to manage investment risk. For example, investment risk models help a portfolio manager understand how much the value of the portfolio is likely to change given a change in a certain risk factor. They also provide insight into the gains and losses the portfolio might reasonably be expected to experience and the frequency with which large losses might occur.

Effective risk management, though, is much more than just applying financial models. It requires the application of judgment and experience not only to know how to use the models appropriately but also to appreciate the strengths and limitations of the models and to know when to supplement or substitute one model with another model or approach.

Financial markets operate more or less continuously, and new prices are constantly being generated. As a result, there is a large amount of data on market risk and a lot of collective experience dealing with this risk, making market risk one of the easier financial risks to analyze. Still, market risk is not an easy risk to capture. Although a portfolio's exposures can be identified with some certainty, the potential losses that could arise from those exposures are unknown. The data used to estimate potential losses are generated from past prices and rates, not the ones to come. Risk management models allow the experienced risk manager to blend that historical data with their own forward-looking judgment, providing a framework within which to test that judgment.

We first lay a foundation for understanding value at risk, discuss three primary approaches to estimating value at risk, and cover the primary advantages and limitations as well as extensions of value at risk. We then address the sensitivity measures used for equities, fixed-income securities, and options and also cover historical and hypothetical scenario risk measures. Next, we discuss the use of constraints in risk management, such as risk budgeting, position limits, scenario limits, stop-loss limits, and capital allocation as risk management tools. Lastly, we describe various applications and limitations of risk measures as used by different types of market participants and summarize our discussion.

Understanding Value at Risk

Value at risk (VaR) was developed in the late 1980s, and over the next decade, it emerged as one of the most important risk measures in global financial markets.

Value at Risk: Formal Definition

Value at risk is the minimum loss that would be expected a certain percentage of the time over a certain period of time given the assumed market conditions. It can be expressed in either currency units or as a percentage of portfolio value. Although this statement is an accurate definition of VaR, it does not provide sufficient clarity to fully comprehend the concept. To better understand what VaR means, let us work with an example. Consider the statement:

> *The 5% VaR of a portfolio is €2.2 million over a one-day period.*

The following three points are important in understanding the concept of VaR:

- VaR can be measured in either currency units (in this example, the euro) or in percentage terms. In this example, if the portfolio value is €400 million, the VaR expressed in percentage terms would be 0.55% (€2.2 million/€400 million = 0.0055).
- VaR is a *minimum* loss. This point cannot be emphasized enough. VaR is often mistakenly assumed to represent *how much one can lose*. If the question is, "how much can one lose?" there is only one answer: *the entire portfolio*. In a €400 million portfolio, assuming no leverage, the most one can lose is €400 million.
- A VaR statement references a time horizon: losses that would be expected to occur over a given period of time. In this example, that period of time is one day. (If VaR is measured on a daily basis, and a typical month has 20–22 business days, then 5% of the days equates to about one day per month.)

These are the explicit elements of a VaR statement: the *frequency* of losses of a given *minimum magnitude* expressed either in *currency* or *percentage* terms. Thus, the VaR statement can be rephrased as follows: A loss of at least €2.2 million would be expected to occur about once every month.

A 5% VaR is often expressed as its complement—a 95% level of confidence. In this reading, we will typically refer to the notion as a 5% VaR, but we should be mindful that it does imply a 95% level of confidence.

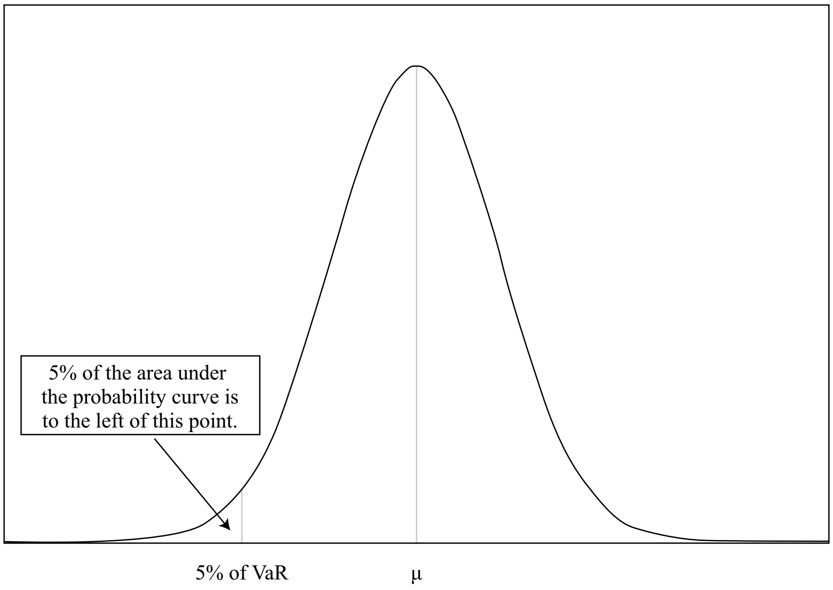
Using the example given, it is correct to say any of the following:

- €2.2 million is the minimum loss we would expect 5% of the time.
- 5% of the time, losses would be at least €2.2 million.
- We would expect a loss of no more than €2.2 million 95% of the time.

The last sentence is sometimes mistakenly phrased as "95% of the time we would expect to lose less than €2.2 million," but this statement could be taken to mean that 95% of the time we would incur losses, although those losses would be less than €2.2 million. In fact, a large percentage of the time we will make money.

Exhibit 1 illustrates the concept of VaR using the 5% case. It depicts a probability distribution of returns from a hypothetical portfolio. The distribution chosen is the familiar normal distribution, known sometimes as the bell curve, but that distribution is only one curve that might be used. In fact, there are compelling arguments that the normal distribution is not the right one to use for financial market returns. We discuss these arguments later.

Exhibit 1: Illustration of 5% VaR in the Context of a Probability Distribution



Note that the distribution in Exhibit 1 is centered on the value μ. [The symbol μ (Greek: *mu*) is a common symbol used to represent an expected value.] Near the left tail of the distribution is the notation "5% VaR," indicating that 5% of the area under the curve is to the left of the point of the VaR (i.e., the probability of observing a value less than the VaR is 5%).

Thus, it is apparent that VaR is simply a point on the probability distribution of profits or returns from a portfolio. Given the characteristics of the normal distribution, a 5% VaR is equivalent to the point on the distribution that is 1.65 standard deviations below the expected value. Although the concept of VaR can be easily visualized in this manner, actually measuring the VaR is a challenge.

Before we take on that challenge, however, note that there is no formal requirement that VaR be measured at a 5% threshold. It is also common to use a 1% threshold (2.33 standard deviations from the expected value), and some investment managers use a one standard deviation movement (equal to a 16% VaR)—both assuming a normal distribution. There is no definitive rule for what VaR cutoff should be used. A specification with a higher confidence level will produce a higher VaR. It is up to the decision maker to choose an appropriate level.

VaR and Standard Deviations

The 16% VaR relates to a one standard deviation move as follows: In a normal distribution, 50% of the outcomes are to the right of the expected value and 50% are to the left. A one standard deviation interval implies that 68% of the outcomes lie within one standard deviation of the expected value; thus, 34% of the outcomes lie one standard deviation to the left of the expected value and 34% of the outcomes one standard deviation to the right. Adding the 50% of the outcomes that lie to the right of the expected value to the 34% of the outcomes that lie one standard deviation below the expected value means that 84% of all outcomes lie to the right of the point that is one standard deviation to the left of the expected value. Therefore, 16% of all outcomes lie below this point. Thus, a one standard deviation movement is equivalent to a 16% VaR (or an 84% level of confidence).

Just as there is no formal requirement that VaR be measured at a 5% cutoff, there is also no formal requirement that VaR be measured using a daily loss estimate. One could reasonably measure VaR on a weekly, bi-weekly, monthly, quarterly, semiannually, or annual basis. Choosing the VaR threshold and the time horizon are examples of why VaR is not a precise measure but in fact entails considerable judgment.

We should also reiterate that VaR can be expressed as a rate of return or in monetary terms. It is typically easier to process the data necessary to estimate VaR in terms of returns, but VaR is most frequently expressed in terms of profits or losses. This point will become clearer as we work through examples.

EXAMPLE 1

Definition of VaR

1. Given a VaR of $12.5 million at 5% for one month, which of the following statements is correct?

   A. There is a 5% chance of losing $12.5 million over one month.

   B. There is a 95% chance that the expected loss over the next month is less than $12.5 million.

   C. The minimum loss that would be expected to occur over one month 5% of the time is $12.5 million.

Solution:

C is correct because it is the only statement that accurately expresses the VaR. A is incorrect because VaR does not give the likelihood of losing a specific amount. B is incorrect because VaR is not an expected loss; rather, it is a minimum loss.

2. Which of the following statements is **not** correct?

   A. A 1% VaR implies a downward move of 1%.

   B. A one standard deviation downward move is equivalent to a 16% VaR.

   C. A 5% VaR implies a move of 1.65 standard deviations less than the expected value.

Solution:

A is correct. A 1% VaR (99% confidence) is the point on the distribution 2.33 standard deviations below the expected value. Answers B and C correctly describe a 16% and 5% VaR, respectively.

To this point, we have given only the conceptual definition of VaR. Defining something is one thing; measuring it can be quite challenging. Such is the case for VaR.

2 ESTIMATING VAR

Three methods are typically used to estimate VaR: the parametric (variance–covariance) method, the historical simulation method, and the Monte Carlo simulation method. Each of these will be discussed in turn.

The first step of every VaR calculation, regardless of the VaR method used, is to convert the set of holdings in the portfolio into a set of exposures to **risk factors**, a process called **risk decomposition**. In some instances, this process can be very simple: An equity security can be the risk factor itself. In other instances, the process can be highly complex. For example, a convertible bond issued by a foreign entity has both currency and equity risk factors as well as exposures to multiple points on a yield curve of a given credit quality. Fixed-income instruments and derivatives products often contain distinct risk exposures that require decomposition in order to accurately capture their loss potential.

The second step of VaR estimation requires gathering a data history for each of the risk factors in the VaR model. The three methods use different approaches to specifying these inputs, which will be discussed in the following sections. We will see that the parametric and Monte Carlo methods do not formally require a data history. They require only that the user enter estimates of certain parameters into the computational procedure (expected return, standard deviation, and for some models, skewness and kurtosis). One of the most common sources for estimating parameter inputs for any financial model is historical data, but the user could substitute estimates based on judgement or alternative forecasting models. Indeed, shortly we will override some historical estimates with our own judgement. Nonetheless, the collection of a data history is typically used at least as a starting point in the parametric and Monte Carlo methods, and it is absolutely required for the historical simulation method.

The third step of each method is where the differences between the three VaR methods are most apparent: how each method uses the data to make an estimate of the VaR.

Although most portfolios contain a large number of individual securities and other assets, we will use a two-asset portfolio to illustrate the three VaR methods. Using a limited number of assets permits us to closely observe the essential elements of the VaR estimation procedure without getting mired in the complex mathematics required to accommodate a large number of assets. The objective is to understand the concept of VaR, be aware of how it is estimated, know how it is used, appreciate the benefits of VaR, and be attentive to its limitations. We can achieve these objectives by keeping the portfolio fairly simple.

Our example portfolio has a market value of $150 million and consists of two ETFs—SPDR S&P 500 ETF (SPY), representing the US equity exposure, and SPDR Portfolio Long-Term Corporate Bond ETF (SPLB), representing a corporate bond exposure. We will allocate 80% of the portfolio to SPY and 20% of the portfolio to SPLB. For the sake of simplicity, the two securities will represent the risk factors and the return history of each ETF will serve as the risk factor history used in the VaR model. We have collected a set of two years of daily total return data, reflecting both capital appreciation and dividends on each ETF. The period used for this historical data set is called the **lookback period**. The question of exactly how much data are required to be a representative data set is a complex question that is common to all estimation problems in economics and finance. We will discuss some of the issues on this matter later in this reading.

Exhibit 2 provides statistical summary information based on the four years of daily data in the lookback period, covering the period of 1 July 2015 through 28 June 2019.

Exhibit 2: Statistical Estimates from Daily Return Data, 1 July 2015–28 June 2019

	Daily		Annualized	
	Average Return	Standard Deviation	Average Return	Standard Deviation
SPY	0.047%	0.86%	12.51%	13.64%
SPLB	0.031%	0.49%	8.03%	7.73%

Note: The correlation of SPLB and SPY = −0.0607.

SPY produced an annualized average return of about 12.5% with a standard deviation of 13.6%, significantly different from the long-term historical performance of the S&P 500 Index of approximately 10.5% average return and 20% standard deviation. SPLB produced an annualized average return of 8% with a standard deviation of about 7.7%. These numbers compare with an average annual return for long-term corporate bonds of slightly more than 6% and a standard deviation of about 8.5% (historical data are drawn from Malkiel 2007). Although the average return of SPLB in the last four years was higher than that of the overall long-term corporate bond sector, the standard deviations were similar.

The risk and return parameters for each risk factor in Exhibit 2 illustrate how one might collect historical data. It is necessary, however, to critically assess the data and apply judgment to modify the inputs if the lookback period is not representative of the expected performance of the securities (or risk factors) going forward. Exercising our judgment, and believing that we have no information to suggest that future performance will deviate from the long-run historical performance, we adjust our inputs and use returns of 10.5% for SPY and 6% for SPLB, with standard deviations of 20% for SPY and 8.5% for SPLB. These adjustments align the inputs more closely with the long-run historical performance of each sector. In practice, users will want to use estimates they believe are reflective of current expectations, though clearly one user's estimates could differ widely from another's.

Although the returns and standard deviations experienced over the lookback period have been adjusted to more closely align with long-run historical experience, we will use a correlation estimate approximately equal to the observed correlation over our lookback period. We are assuming that the recent historical relationship of equity and fixed-income returns is a reasonable assumption moving forward. To keep the numbers simple, we round the observed correlation of −0.0607 to −0.06.

Exhibit 3 illustrates our input assumptions for the VaR estimations.

Exhibit 3: Input Assumptions, 1 July 2015–28 June 2019

		Annualized	
	Allocation	Return	Standard Deviation
SPY	80%	10.5%	20.0%
SPLB	20%	6.0%	8.5%

Note: The correlation of SPLB and SPY = −0.06.

3 THE PARAMETRIC METHOD OF VAR ESTIMATION

- ☐ compare the parametric (variance–covariance), historical simulation, and Monte Carlo simulation methods for estimating VaR
- ☐ estimate and interpret VaR under the parametric, historical simulation, and Monte Carlo simulation methods

The **parametric method** of estimating VaR is sometimes referred to as the analytical method and sometimes the variance–covariance method. The parametric method begins, as does each method, with a risk decomposition of the portfolio holdings. It typically assumes that the return distributions for the risk factors in the portfolio are normal. It then uses the expected return and standard deviation of return for each risk factor to estimate the VaR.

Note that we said that this method *typically* uses the normal distribution. Indeed, that is the common case in practice, but there is no formal requirement that the normal distribution be used. The normal distribution conveniently requires only two parameters—the expected value and standard deviation—to encompass everything there is to know about it. If other distributions are used, additional parameters of the distribution, such as skewness and kurtosis, would be required. We will limit the presentation here to the normal distribution, but be aware that other, more accurately representative distributions could be used but would add complexity to the VaR estimation process.

Recall that in defining VaR, we identified a VaR threshold—a point in the left tail of the distribution, typically either the 5% left tail, the 1% left tail, or a one standard deviation move (16%). If the portfolio is characterized by normally distributed returns and the expected value and standard deviation are known, it is a simple matter to identify any point on the distribution. A normal distribution with expected value μ and standard deviation σ can be converted to a standard normal distribution, which is a special case of the normal distribution in which the expected value is zero and the standard deviation is one. A standard normal distribution is also known as a *z*-distribution. If we have observed a return *R* from a normal distribution, we can convert to its equivalent *z*-distribution value by the transformation:

$$z = \frac{R - \mu}{\sigma}.$$

In a standard normal (*z*) distribution, a 5% VaR is 1.65 standard deviations below the expected value of zero. A 1% VaR is 2.33 standard deviations below the expected value of zero. A 16% VaR is one standard deviation below the expected value of zero. Thus, in our example, for a 5% VaR, we wish to know the return that is 1.65 standard deviations to the left of the expected return.

To estimate this VaR, we need the expected return and volatility of the portfolio. The expected return is estimated from the following equation:

$$E(R_p) = w_{SPY}E(R_{SPY}) + w_{SPLB}E(R_{SPLB}), \tag{1}$$

where the expected return of the portfolio, $E(R_p)$, is equal to the portfolio weights of SPY (w_{SPY}) and SPLB (w_{SPLB}) multiplied by the expected return of each asset, $E(R_{SPY})$ and $E(R_{SPLB})$.

The volatility of the portfolio, σ_p, is estimated from the following equation:

$$\sigma_p = \sqrt{w_{SPY}^2\sigma_{SPY}^2 + w_{SPLB}^2\sigma_{SPLB}^2 + 2\,w_{SPY}w_{SPLB}\rho_{SPY,SPLB}\sigma_{SPY}\sigma_{SPLB}}, \tag{2}$$

where σ_{SPY} and σ_{SPLB} are the standard deviations (volatilities) of SPY and SPLB, respectively; $\rho_{SPY,SPLB}$ is the correlation between the returns on SPY and SPLB, respectively; and $\rho_{SPY,SPLB}\sigma_{SPY}\sigma_{SPLB}$ is the covariance between SPY and SPLB.

Recall that we estimated these parameters from the historical data, with some modifications to make them more consistent with long-run values. The formal calculations for our portfolio based on these adjusted estimates are as follows:

$$E\left(R_p\right) = 0.8\,(0.105) + 0.2\,(0.06) = 0.096000$$

$$\sigma_p = \sqrt{(0.8)^2(0.2)^2 + (0.2)^2(0.085)^2 + 2\,(0.8)\,(0.2)\,(-0.06)\,(0.2)\,(0.085)}$$

$$=0.159883.$$

Thus, our portfolio, consisting of an 80% position in SPY and a 20% position in SPLB, is estimated to have an expected return of 9.6% and a volatility of approximately 15.99%.

But these inputs are based on annual returns. If we want a one-day VaR, we should adjust the expected returns and volatilities to their daily counterparts. Assuming 250 trading days in a year, the expected return is adjusted by dividing by 250 and the standard deviation is adjusted by dividing by the square root of 250. (Note that the variance is converted by dividing by time, 250 days; thus, the standard deviation must be adjusted by using the square root of time, 250 days.) Thus, the daily expected return and volatility are

$$E\left(R_p\right) = \frac{0.096}{250} = 0.000384 \tag{3}$$

and

$$\sigma_p = \frac{0.159883}{\sqrt{250}} = 0.010112. \tag{4}$$

It is important to note that we have assumed that the statistical properties of the return distribution are constant across the year. Earlier, we annualized the daily data in Exhibit 2 in order to see how our estimates compared with long-term estimates. We made some modest adjustments to the annualized data and then, in Equations 3 and 4, returned to using daily data. To estimate an annual VaR, we would need to use annual data, but we would need a longer lookback period in order to have sufficient data points.

It is important to note that we cannot estimate a daily VaR and annualize it to arrive at an annual VaR estimate. First, to assume that a daily distribution of returns can be extrapolated to an annual distribution is a bold assumption. Second, annualizing the daily VaR is not the same as adjusting the expected return and the standard deviation to annual numbers and then calculating the annual VaR. The expected return is annualized by multiplying the daily return by 250, and the standard deviation is annualized by multiplying the daily standard deviation by the square root of 250. Thus, we can annualize the data and estimate an annual VaR, but we cannot estimate a daily VaR and annualize it without assuming a zero expected return.

Having calculated the daily expected return and volatility, the parametric VaR is now easily obtained. With the distribution centered at the expected return of 0.0384% and a one standard deviation move equal to 0.996%, a 5% VaR is obtained by identifying the point on the distribution that lies 1.65 standard deviations to the left of the mean. It is now easy to see why parametric VaR is so named: The expected values, standard deviations, and covariances are the *parameters* of the distributions.

The following step-by-step procedure shows how the VaR is derived:

$\{[E(Rp) - 1.65\sigma_p](-1)\}(\$150{,}000{,}000)$

Step 1 Multiply the portfolio standard deviation by 1.65.

$0.010112 \times 1.65 = 0.016685$

Step 2 Subtract the answer obtained in Step 1 from the expected return.

$0.000384 - 0.016685 = -0.016301$

Step 3 Because VaR is expressed as an absolute number (despite representing an expected loss), change the sign of the value obtained in Step 2.

Change −0.016301 to 0.016301

Step 4 Multiply the result in Step 3 by the value of the portfolio.

$\$150{,}000{,}000 \times 0.016301 = \$2{,}445{,}150$

Thus, using the parametric method, our estimate of VaR is $2,445,150, meaning that on 5% of trading days the portfolio would be expected to incur a loss of at least $2,445,150. Note that asset managers may stop at Step 3 because at that point the measure is expressed as a percentage of the value of the portfolio, which is the unit this group more commonly uses.

EXAMPLE 2

Parametric VaR

1. The parameters of normal distribution required to estimate parametric VaR are:

   A. expected value and standard deviation.

   B. skewness and kurtosis.

   C. standard deviation and skewness.

Solution:

A is correct. The parameters of a normal distribution are the expected value and standard deviation. Skewness, as mentioned in B and C, and kurtosis, as mentioned in B, are characteristics used to describe a *non*-normal distribution.

2. Assuming a daily expected return of 0.0384% and daily standard deviation of 1.0112% (as in the example in the text), which of the following is *closest* to the 1% VaR for a $150 million portfolio? Express your answer in dollars.

   A. $3.5 million

   B. $2.4 million

   C. $1.4 million

Solution:

A is correct and is obtained as follows:

Step 1 $2.33 \times 0.010112 = 0.023561$

Step 2 $0.000384 - 0.023561 = -0.023177$

Step 3 Convert −0.023177 to 0.023177

Step 4 0.023177 × $150 million = $3,476,550

B is the estimated VaR at a 5% threshold, and C is the estimated VaR using a one standard deviation threshold.

3. Assuming a daily expected return of 0.0384% and daily standard deviation of 1.0112% (as in the example in the text), the daily 5% parametric VaR is $2,445,150. Rounding the VaR to $2.4 million, which of the following values is *closest* to the annual 5% parametric VaR? Express your answer in dollars.

   A. $38 million

   B. $25 million

   C. $600 million

Solution:

B is correct. It is found by annualizing the daily return and standard deviation and using these figures in the calculation. The annual return and standard deviation are, respectively, 0.096000 (0.000384 × 250) and 0.159885 (0.010112 × $\sqrt{250}$).

Step 1 0.159885 × 1.65 = 0.263810

Step 2 0.096000 − 0.263810 = −0.167810

Step 3 Convert −0.167810 to 0.167810

Step 4 0.167810 × $150 million = $25,171,500

A incorrectly multiplies the daily VaR by the square root of the number of trading days in a year ($\sqrt{250}$), and C incorrectly multiplies the daily VaR by the approximate number of trading days in a year (250). Neither A nor C make the appropriate adjustment to annualize the standard deviation.

To recap, we see that the parametric VaR method generally makes the assumption that the distribution of returns on the risk factors is normal. Under that assumption, all of the information about a normal distribution is contained in the expected value and standard deviation. Therefore, finding the 5% VaR requires only that we locate the point in the distribution beyond which 5% of the outcomes occur. Although normality is the general assumption of the parametric method, it is not an absolute requirement. Other distributions could be accommodated by incorporating skewness and kurtosis, the third and fourth parameters of the distribution, but that added complexity is not needed to demonstrate the general approach to parametric VaR and is rarely done in practice.

The major advantage of the parametric method is its simplicity and straightforwardness. The assumption of the normal distribution allows us to easily estimate the parameters using historical data, although judgment is required to adjust the parameters when the historical data may be misleading. The parametric method is best used in situations in which one is confident that the normal distribution can be applied as a reasonable approximation of the true distribution and the parameter estimates are reliable or can be turned into reliable estimates by suitable adjustments. It is important to understand that VaR under the parametric method is very sensitive to the parameter estimates, especially the covariances.

One of the major weaknesses of the parametric method is that it can be difficult to use when the investment portfolio contains options. When options are exercised, they pay off linearly with the underlying; however, if never exercised, an option loses 100% of its value. This characteristic leads to a truncated, non-normal distribution that does not lend itself well to the parametric method. But some adjustments can render options more responsive to the parametric method. These adjustments are helpful but not perfect, limiting the usefulness of the parametric method when options are in the portfolio. Additionally, although the expected return and volatility of the underlying fixed income or equity security may be stable over the life of the option, the distribution of the option changes continuously as the value of the underlying, the volatility of the underlying, and the time to expiration all change.

4 THE HISTORICAL SIMULATION METHOD OF VAR ESTIMATION

- ☐ compare the parametric (variance–covariance), historical simulation, and Monte Carlo simulation methods for estimating VaR
- ☐ estimate and interpret VaR under the parametric, historical simulation, and Monte Carlo simulation methods

The **historical simulation method** of VaR uses the *current* portfolio and reprices it using the actual *historical* changes in the key factors experienced during the lookback period. We begin, as with the parametric method, by decomposing the portfolio into risk factors and gathering the historical returns of each risk factor from the chosen lookback period. Unlike the parametric method, however, we do not characterize the distribution using estimates of the mean return, the standard deviation, or the correlations among the risk factors in the portfolio. Instead, we reprice the current portfolio given the returns that occurred on each day of the historical lookback period and sort the results from largest loss to greatest gain. To estimate a one-day VaR at a 5% confidence interval, we choose the point on the resulting distribution beyond which 5% of the outcomes result in larger losses.

Illustrating this point using a full four years of daily observations would be tedious and consume a great deal of space, so we will condense the process quite a bit and then extrapolate the methodology. Exhibit 4 shows the daily returns on the SPY, the SPLB, and our 80% SPY/20% SPLB portfolio over the first five days of our historical data set. Please note that fixed weights are assumed for all days. Neither historical simulation nor Monte Carlo simulation is intended to be a replication of sequences of prices. They are intended to create a sample of one-day returns for a portfolio of given weights.

Exhibit 4: First Five Days of Historical Returns on the SPY/SPLB Portfolio Using the 1 July 2015–28 June 2019 Data

Day	SPY Return	SPLB Return	Portfolio Return
1	0.80%	−0.53%	0.53%
2	−0.09%	0.45%	0.02%
3	−0.28%	1.47%	0.07%
4	−0.63%	0.28%	0.56%
5	−1.68%	−0.23%	−1.39%

Notes: The Day 1 portfolio return is obtained by multiplying each holding (SPY, SPLB) by its respective weight in the portfolio (80%/20%) and adding the two results together: 0.80(0.008) + 0.20(−0.0053). Although Exhibit 4 shows only five days of returns, we would, of course, use all of the data at our disposal that is reasonably representative of possible future outcomes.

The historical simulation VaR extracts the portfolio return that lies at the appropriate confidence interval along the distribution. Using Excel's "=percentile(x,y)" function, we calculated the following historical simulation VaRs for our sample portfolio:

- 1% VaR (99% confidence) $2,643,196
- 5% VaR (95% confidence) $1,622,272
- 16% VaR (84% confidence) $880,221

Now, it will be interesting to compare this result with the parametric VaR estimates. Exhibit 5 shows the results side-by-side with the parameters used. The historical simulation method does not directly use these parameters but uses the data itself, and these numbers are the parameters implied by the data itself.

Exhibit 5: Comparison of Historical and Parametric VaR Estimates Using 1 July 2015–28 June 2019 Data

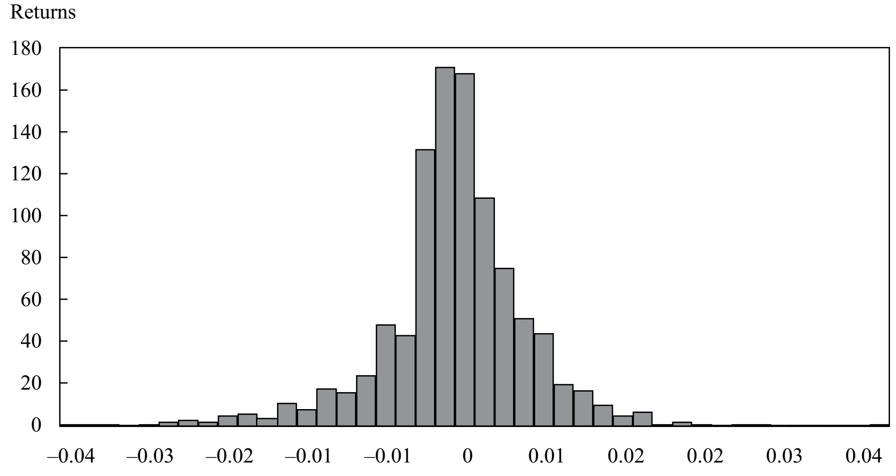
	Historical Simulation Method	Parametric Method
1% VaR	$2,643,196	$3,476,550
5% VaR	$1,622,272	$2,445,150
16% VaR	$880,221	$1,459,200

	Average Return	Standard Deviation	Average Return	Standard Deviation
SPY	12.51%	13.64%	10.50%	20.00%
SPLB	8.03%	7.73%	6.00%	8.50%
Correlation of SPY and SPLB	−0.061		−0.06	

The historical simulation VaRs are much smaller, and the differences stem primarily from the adjustments we made to the historical parameters. We adjusted the volatility and the average return estimates of SPY to more closely reflect the historical norms and slightly raised the volatility of SPLB. Recall, in particular, that our factor history for the S&P 500 exhibited abnormally low volatility relative to the long-run experience.

Additionally, our calculations using the historical simulation method were not constrained by the assumption of a normal distribution as was the case with the parametric method. Exhibit 6 is a histogram of the portfolio returns used in the historical simulation results, overlaid with a normal distribution.

Exhibit 6: Histogram of Historical Portfolio Returns (80% SPY and 20% SPLB) Using 1 July 2015–28 June 2019 Data



As can be seen, the resulting distribution under the historical simulation method is a departure from a normal distribution. This point again highlights the importance of understanding the underlying assumptions of any VaR model.

There is *no single right way* of estimating VaR. Each method provides an estimate of VaR and is highly sensitive to the input parameters, and similar to many estimation models, they will disagree.

Both the parametric and historical simulation methods in their most basic forms have the limitation that, as with most samples, all observations are weighted equally. The historical simulation method can adjust for this problem, however, by using a weighting methodology that gives more weight to more recent observations and less weight to more distant observations.

The primary advantage of the historical simulation method compared with the parametric method is that the historical simulation method estimates VaR based on what actually happened, so it cannot be dismissed as introducing impossible outcomes. Yet, therein also lies the primary weakness of the historical simulation method: There can be no certainty that a historical event will re-occur or that it would occur in the same manner or with the same likelihood as represented by the historical data. If one uses a relatively short historical data set, such as from January 1987 through December 1988 (a period encompassing the "Black Monday" of 19 October 1987, when stock markets around the world collapsed in a very short time), an occurrence of this magnitude might be projected to occur once every two years, surely an overstatement of its probability. Thus, the historical simulation method is best used when the distribution of returns during the lookback period are expected to be representative of the future.

The historical method is capable of handling the adjustment of one time horizon to another; that is, the information derived from daily data can be extrapolated to estimate an annual VaR, provided the distribution can be assumed to be stationary. In other words, one can convert each daily return to an annual return and then estimate the annual VaR. Although using annual data to estimate an annual VaR is always preferred, that would require a much longer lookback period.

We noted earlier that the parametric method is not well suited for options. Because the historical simulation method captures the returns that actually occurred regardless of the type of financial instrument used, it can accommodate options.

EXAMPLE 3

Historical Simulation VaR

1. Which of the following statements about the historical simulation method of estimating VaR is *most* correct?

   A. A 5% historical simulation VaR is the value that is 5% to the left of the expected value.

   B. A 5% historical simulation VaR is the value that is 1.65 standard deviations to the left of the expected value.

   C. A 5% historical simulation VaR is the fifth percentile, meaning the point on the distribution beyond which 5% of the outcomes result in larger losses.

Solution:

C is correct. In the historical method, the portfolio returns are arrayed lowest to highest and the observation at the fifth percentile (95% of the outcomes are better than this outcome) is the VaR. A is not correct because it draws a point on the distribution relative to the expected value rather than using the 5% of the outcomes that are in the left-most of the distribution. B confuses the parametric and historical methods. In the parametric method, the 5% VaR lies 1.65 standard deviations below the mean.

2. Which of the following is a limitation of the historical simulation method?

   A. The past may not repeat itself.

   B. There is a reliance on the normal distribution.

   C. Estimates of the mean and variance could be biased.

Solution:

A is correct. The historical simulation method estimates VaR based on the historical distribution of the risk factors. B is not correct; the historical simulation method does not rely on any particular distribution because it simply uses whatever distribution applied in the past. C is not correct because the historical distribution does not formally estimate the mean and variance.

5 THE MONTE CARLO SIMULATION METHOD OF VAR ESTIMATION

- ☐ compare the parametric (variance–covariance), historical simulation, and Monte Carlo simulation methods for estimating VaR
- ☐ estimate and interpret VaR under the parametric, historical simulation, and Monte Carlo simulation methods

Monte Carlo simulation is a method of estimating VaR in which the user develops his own assumptions about the statistical characteristics of the distribution and uses those characteristics to generate random outcomes that represent hypothetical returns to a portfolio with the specified characteristics. This method is widely used in the sciences to estimate the statistical distribution of scientific phenomena and has many applications in business and finance. For example, a corporation considering the investment of a large amount of capital in a new project with many uncertain variables could simulate the possible values of these variables and thus gain an understanding of the distribution of the possible returns from this investment. Or, complex options can often be priced by simulating outcomes of the underlying, determining the payoffs of the option, and then averaging the option payoffs and discounting that value back to the present. The reference to the famous Mediterranean casino city allegedly came from an observation made by a scientist that the method is similar to tossing dice at a casino.

Monte Carlo simulation avoids the complexity inherent in the parametric method when the portfolio has a large number of assets. (A large number of assets makes the parameters of the distribution difficult to extract.) There can be many risk factors, and the interactions among these risk factors can be too complex to specify. Moreover, Monte Carlo simulation does not need to be constrained by the assumption of normal distributions. Rather than attempt to determine the expected return and volatility of a combination of multiple statistical processes, one would simply simulate these processes, tabulate the statistical results of the simulations, and thereby gain a measure of the combined effects of these complex component processes on the overall risk.

Monte Carlo simulation requires the generation of random values of the underlying unknowns. In our example, the unknowns are the returns on the two risk factors, represented by the SPY and SPLB ETFs. We can, of course, assume that the statistical properties of the historical returns—their averages, volatilities, and correlation—are appropriate for use in a simulation, or we can modify those values to conform to what we expect to be relevant for the future. For illustrative purposes here, we will simply use the inputs we used in the parametric method.

Recall that we previously assumed for the sake of simplicity that the two securities represent the risk factors. We now decompose the portfolio holdings into these risk factors. First we simulate the returns of these two risk factors, and then we re-price our exposures to the risk factors under the range of simulated returns, recording the results much as we do in the historical simulation method. We then sort the results in order from worst to best. A 5% Monte Carlo VaR would simply be the fifth percentile of the simulated values instead of the historical values.

Yet, it is not quite that simple. We must first decide how many random values to generate. There is no industry standard. The more values we use, the more reliable our answers are but the more time-consuming the procedure becomes. In addition, we cannot just simulate values of two random variables without accounting for the correlation between the two. For example, if you spin two roulette wheels, you can assume they are independent of each other in much the same manner as are two uncorrelated assets. But most assets have at least a small degree of correlation. In our example, we used the historical correlation of about −0.06. Monte Carlo simulation must take that relationship into account.

For simplicity, this reading will not go into detail on either the mathematical techniques that can account for the correlations among risk factor returns or the specific method used to simulate outcomes given average values and volatilities for each risk factor. Both are beyond the scope of this reading.

For this example, we will use 10,000 simulated returns on SPY and SPLB drawn from a normal distribution. Of course, non-normal distributions can be used—and they commonly are in practice—but we want to keep the illustration simple to facilitate comparisons between methods. Each set of simulated returns combines to produce a sample with the expected returns and volatilities as we specified. In addition, the returns will have the pre-specified correlation of −0.06. Each pair of returns is weighted 80/20 as desired. We generate the 10,000 outcomes, sort them from worst to best, and either select the outcome at the 5th percentile for a 5% VaR, the outcome at the 1st percentile for a 1% VaR, or the outcome at the 16th percentile if we want to evaluate the impact of a one standard deviation move. Using the parameters specified in our example, the simulation returns a distribution from which we can draw the following VaR numbers:

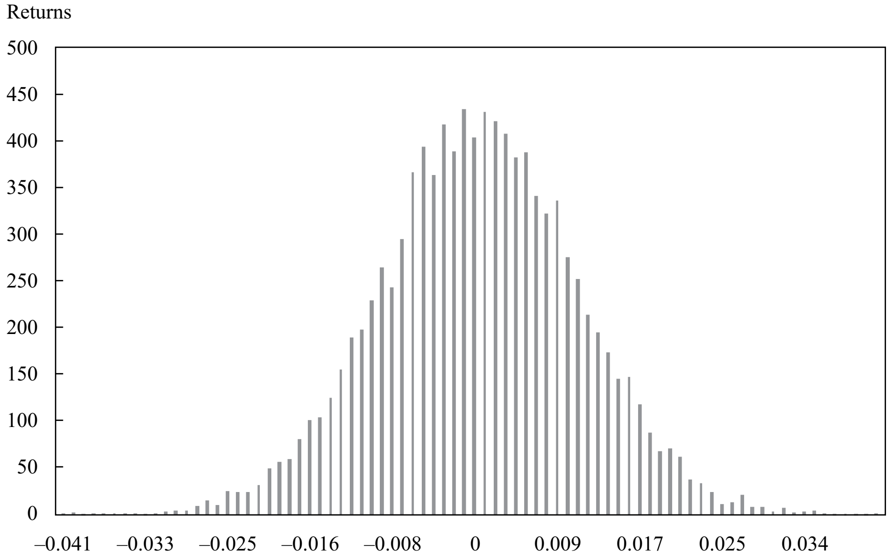
1% VaR = $3,541,035

5% VaR = $2,517,702

16% VaR = $1,524,735

Note that these results are fairly close to VaR under the parametric VaR method, where the 5% VaR was $2,445,150. The slight difference arises from the fact that Monte Carlo simulation only *samples* from a population with certain parameters while the parametric method *assumes* those parameters. A sample of a distribution will not produce statistics that match the parameters precisely except in extremely large sample sizes, much larger than the 10,000 used here. Exhibit 7 displays a histogram of the simulated returns overlaid with a bell curve representing a normal distribution. Note how the simulated returns appear more normally distributed than do the historical values, as illustrated in Exhibit 6. This is because we explicitly assumed a normal distribution when running the simulation to generate the values in our example.

Exhibit 7: Monte Carlo Simulated Returns 80/20 Portfolio of SPY and SPLB



Although we conveniently assumed a normal distribution, one of the advantages of the Monte Carlo method is that it can accommodate virtually *any* distribution. In fact, the flexibility of the Monte Carlo method to handle more complex distributions is its primary attraction. The Monte Carlo and historical simulation methods are much more capable than the parametric method of accurately incorporating the effects of option positions or bond positions with embedded options.

Similar to the historical simulation method, you can scale daily returns to annual returns and extrapolate an estimate of the annual VaR by running a Monte Carlo simulation on these annual returns.

At one time, calculating VaR using the Monte Carlo simulation method was slow, but with the speed of today's computers, it is relatively easy and fast to simulate extremely complex processes for portfolios with thousands of exposures.

EXAMPLE 4

Monte Carlo Simulation VaR

1. When will the Monte Carlo method of estimating VaR produce virtually the same results as the parametric method?

   A. When the Monte Carlo method assumes a non-normal distribution.

   B. When the Monte Carlo method uses the historical return and distribution parameters.

   C. When the parameters and the distribution used in the parametric method are the same as those used in the Monte Carlo method and the Monte Carlo method uses a sufficiently large sample.

Solution:

C is correct. The Monte Carlo method simulates outcomes using whatever distribution is specified by the user. *If* a normal distribution is used *and* a sufficiently large number of simulations are run, the parameters of the Monte Carlo sample will converge with those used in the parametric method and the overall VaR should be very close to that of the parametric method. A is incorrect because the parametric method is not well-adapted to a non-normal distribution. B is incorrect because neither the Monte Carlo method nor the parametric method focuses on historical outcomes.

2. Which of the following is an advantage of the Monte Carlo method?

   A. The VaR is easy to calculate with a simple formula.

   B. It is flexible enough to accommodate many types of distributions.

   C. The number of necessary simulations is determined by the parameters.

Solution:

B is correct. The method can handle any distribution. A is incorrect because Monte Carlo simulation is not a simple formula. C is incorrect; there is no industry-wide agreement as to the necessary number of simulations.

6 ADVANTAGES, LIMITATIONS, AND EXTENSIONS OF VAR

- ☐ describe advantages and limitations of VaR
- ☐ describe extensions of VaR

The concept of VaR is solidly grounded in modern portfolio analysis. Nonetheless, the implementation of VaR, both in the estimation procedure and in the application of the concept, presents a number of advantages and limitations.

Advantages of VaR

The use of VaR as a risk measure has the following advantages:

- *Simple concept*. VaR is relatively easy to understand. Although the methodology is fairly technical, the concept itself is not very difficult. So, decision makers without technical backgrounds should be able to grasp the likelihood of possible losses that might endanger the organization. Reporting that a daily 5% VaR is, for example, €2.2 million allows the user to assess the risk in the context of the capital deployed. If a portfolio is expected to incur losses of a minimum of €2.2 million on 5% of the trading days, about once a month, this information is valuable in the context of the size of the portfolio.
- *Easily communicated concept*. VaR captures a considerable amount of information into a single number. If the recipient of the information fully understands the meaning and limitations of VaR, it can be a very significant and practical piece of information.
- *Provides a basis for risk comparison.* VaR can be useful in comparing risks across asset classes, portfolios, and trading units—giving the risk manager a better picture of which constituents are contributing the least and the most to the overall risk. As such, the risk manager can be better informed as he looks for potential hot spots in the organization. This point will be discussed further in a later section.
- *Facilitates capital allocation decisions*. The ability to compare VaR across trading units or portfolio positions provides management with a benchmark that can be used in capital allocation decisions. A proprietary trading firm, for example, can find that its VaR in equity trading is $20 million and its VaR in fixed-income trading is $10 million. If its equity trading portfolio is not expected to take more risk than its fixed-income trading portfolio, then the equity trading activities are taking too much risk or there is too much capital allocated to equity trading. The firm should either make adjustments to realign its VaR or allocate capital in proportion to the relative risks. If a firm is looking to add a position to a portfolio or change the weights of existing portfolio positions, certain extensions of VaR allow the manager to assess the risk of these changes. This topic will be covered in more detail later.
- *Can be used for performance evaluation*. Risk-adjusted performance measurement requires that return or profit be adjusted by the level of risk taken. VaR can serve as the basis for risk adjustment. Without this adjustment, more profitable units could be perceived as more successful; however, when adjusted by VaR, a less profitable unit that poses less risk of loss may be judged more desirable.
- *Reliability can be verified*. VaR is easily capable of being verified, a process known as backtesting. For example, if the daily VaR is $5 million at 5%, we would expect that on 5% of trading days a loss of at least $5 million would be incurred. To determine whether a VaR estimate is reliable, one can determine over a historical period of time whether losses of at least $5 million were incurred on 5% of trading days, subject to reasonable statistical variation.
- *Widely accepted by regulators*. In the United States, the SEC requires that the risk of derivatives positions be disclosed either in the form of a summary table, by sensitivity analysis (a topic we cover later), or by VaR. Thus, VaRs are frequently found in annual reports of financial firms. Global banking

regulators also encourage banks to use VaR. These regulations require or encourage the use of VaR, but they do not prescribe how it should be implemented, which estimation method to use, or the maximum acceptable VaR.

Limitations of VaR

Despite its many advantages, users of VaR must also understand its limitations. The primary limitations of VaR are the following:

- *Subjectivity*. In spite of the apparent scientific objectivity on which it is based, VaR is actually a rather subjective method. As we saw in the descriptions of the three methods of estimating VaR, there are many decisions to make. At the fundamental level, decisions must be made as to the desired VaR cutoff (5%, 1%, or some other cutoff); over what time horizon the VaR will be measured; and finally, which estimation method will be used. As we have seen here, for each estimation method, there are numerous other discretionary choices to make about inputs, source of data, and so on.
- *Underestimating the frequency of extreme events*. In particular, use of the normal distribution in the parametric method and sometimes in the Monte Carlo method commonly underestimates the likelihood of extreme events that occur in the left tail of the distribution. In other words, there are often more extreme adverse events, called "left-tail events," than would be expected under a normal distribution. As mentioned previously, there is no particular requirement that one use the normal distribution. The historical simulation method uses whatever distribution the data produce. We chose to illustrate the Monte Carlo method with a normal distribution, and it is virtually always used in the parametric method. Nonetheless, the tendency to favor the normal distribution and other simple and symmetrical distributions often leads to an understatement of the frequency of left-tail events.
- *Failure to take into account liquidity*. If some assets in a portfolio are relatively illiquid, VaR could be understated, even under normal market conditions. Additionally, liquidity squeezes are frequently associated with tail events and major market downturns, thereby exacerbating the risk. Although illiquidity in times of stress is a general problem that affects virtually all of a firm's financial decisions, reliance on VaR in non-normal market conditions will lead the user to underestimate the magnitude of potential losses.
- *Sensitivity to correlation risk*. Correlation risk is the risk that during times of extreme market stress, correlations among all assets tend to rise significantly. Thus, markets that provide a reasonable degree of diversification under normal conditions tend to decline together under stressed market conditions, thereby no longer providing diversification.
- *Vulnerability to trending or volatility regimes.* A portfolio might remain under its VaR limit every day but lose an amount approaching this limit each day. Under such circumstances, the portfolio could accumulate substantial losses without technically breaching the VaR constraint. Also, during periods of low volatility, VaR will appear quite low, underestimating the losses that could occur when the environment returns to a normal level of volatility.
- *Misunderstanding the meaning of VaR*. VaR is not a worst-case scenario. Losses can and will exceed VaR.

- *Oversimplification*. Although we noted that VaR is an easily communicated concept, it can also oversimplify the picture. And although VaR does indeed consolidate a considerable amount of information into a single number, that number should be interpreted with caution and an awareness of the other limitations as well as supported by additional risk measures.
- *Disregard of right-tail events*. VaR focuses so heavily on the left tail (the losses) that the right tail (potential gains) are often ignored. By examining both tails of the distribution, the user can get a better appreciation of the overall risk–reward trade-off, which is often missed by concentrating only on VaR.

These limitations are not unique to VaR; they apply equally to any technique or measure used to quantify the expected rewards and risks of investing.

EXAMPLE 5

Advantages and Limitations of VaR

1. Which of the following is **not** an advantage of VaR?

   A. It is a simple concept to communicate.

   B. There is widespread agreement on how to calculate it.

   C. It can be used to compare risk across portfolios or trading units.

Solution:

B is correct. There is no consensus on how to calculate VaR. A and C are both advantages of VaR, as we noted that VaR is fairly simple to communicate and it can show the contribution of each unit to the overall VaR.

2. Which of the following is a limitation of VaR?

   A. It requires the use of the normal distribution.

   B. The maximum VaR is prescribed by federal securities regulators.

   C. It focuses exclusively on potential losses, without considering potential gains.

Solution:

C is correct. VaR deals exclusively with left-tail or adverse events. A is wrong because although parametric VaR does generally use the normal distribution, the historical simulation method uses whatever distribution occurred in the past and Monte Carlo simulation uses whatever distribution the user chooses. B is incorrect because regulators do not specify maximum VaRs, although they may encourage and require that the measure be used.

Extensions of VaR

Clearly no single risk model can answer all of the relevant questions a risk manager may have. As a result, VaR has laid a foundation for a number of variations, each of which provides additional information.

As discussed previously, VaR is a minimum loss and is typically expressed as the minimum loss that can be expected to occur 5% of the time. An important and related measure can determine the average loss that would be incurred if the VaR cutoff is exceeded. This measure is sometimes referred to as the **conditional VaR (CVaR)**, although it is not technically a VaR measure. It is the average loss conditional on exceeding the VaR cutoff. So, VaR answers the question, "What is the minimum loss I can expect at a certain confidence?" And CVaR answers the question, "How much can I expect to lose if VaR is exceeded?" CVaR is also sometimes referred to as the **expected tail loss** or **expected shortfall**. CVaR is best derived using the historical simulation and Monte Carlo methods, in which one can observe all of the returns throughout the distribution and calculate the average of the losses beyond the VaR cutoff. The parametric method uses a continuous distribution, so obtaining the average loss beyond the VaR cutoff would require a level of mathematics beyond the scope of this reading.

Using our earlier example, in the historical simulation method, our sample of 500 historical returns was sorted from lowest to highest and the 5% VaR was $1,622,272. With 1,006 returns in the sample, 50 observations (5% of 1,006) lie below the VaR estimate. The average of these losses is $2,668,389. Thus, when the VaR is exceeded, we would expect an average loss of about $2.7 million.

For the Monte Carlo method, we generated 10,000 random values and obtained a 5% VaR of $2,517,705. Given 10,000 random values, 500 observations are in the lowest 5% of the VaR distribution. The CVaR using the Monte Carlo method would be the average of the 500 lowest values, which is $4,397,756.

Note that once again, the CVaR derived using the historical simulation method is lower than the CVaR derived using the Monte Carlo method. As explained earlier, this result can largely be attributed to the lower volatility of the S&P 500 component in the historical data series.

Beyond assessing tail loss, a risk manager often wants to know how the portfolio VaR will change if a position size is changed relative to the remaining positions. This effect can be captured by a concept called **incremental VaR (IVaR)**. Using our example, suppose the portfolio manager is contemplating increasing the risk by increasing the investment in SPY to 90% of the portfolio. We recalculate the VaR under the proposed allocation, and the incremental VaR is the difference between the "before" and "after" VaR. As an example, using the parametric method, the VaR would be expected to increase from $2,445,150 to $2,752,500; thus, the IVaR for the 5% case would be $307,350. Or, the portfolio manager might wish to add a new asset, thereby reducing the exposure to the existing assets. The risk manager would calculate the VaR under the assumption that the change is made, and then the difference between the new VaR and the old VaR is the IVaR. This measure is useful because it reflects the effect of an anticipated change on the VaR. The risk manager could find that the new VaR will be unacceptably high or that it has possibly even decreased.

A related concept is called **marginal VaR (MVaR)**. It is conceptually similar to incremental VaR in that it reflects the effect of an anticipated change in the portfolio, but it uses formulas derived from calculus to reflect the effect of a very small change in the position. Some people interpret MVaR as a change in the VaR for a $1 or 1% change in the position, although that is not strictly correct. Nonetheless, this interpretation is a reasonable approximation of the concept behind marginal VaR, which is to reflect the impact of a small change. In a diversified portfolio, marginal VaR may be used to determine the contribution of each asset to the overall VaR; the marginal VaRs for all positions may be proportionately weighted to sum to the total VaR.

Both incremental and marginal VaR address the question of what impact a change in the portfolio holdings might have on the total VaR of the portfolio. Both take into account the potential diversifying effects of various positions or subportfolios, and thus they both can be useful in evaluating the potential effect of a trade before the trade is done.

Another related measure is ***ex ante* tracking error**, also known as **relative VaR**, which is a measure of the degree to which the performance of a given investment portfolio might deviate from its benchmark. It is computed using any of the standard

VaR models, described earlier, but the portfolio to which VaR is applied contains the portfolio's holdings *minus* the holdings in the specified benchmark. In other words, the benchmark's holdings, weighted in proportion to the value of the subject portfolio, are entered into the VaR modeling process as short positions. VaR for this measure is typically expressed as a one standard deviation annualized measure. If the portfolio is a perfect match to the benchmark, *ex ante* tracking error will be at or near zero. The more the portfolio differs from the benchmark, the larger the *ex ante* tracking error will be.

EXAMPLE 6

Extensions of VaR

1. Conditional VaR measures the:

   A. VaR over all possible losses.

   B. VaR under normal market conditions.

   C. average loss, given that VaR is exceeded.

Solution:

C is correct. Conditional VaR is the average loss conditional on exceeding the VaR. A is not correct because CVaR is not concerned with losses that do not exceed the VaR threshold, and B is incorrect because VaR does not distinguish between normal and non-normal markets.

2. Which of the following correctly identifies incremental VaR?

   A. The change in VaR from increasing a position in an asset.

   B. The increase in VaR that might occur during extremely volatile markets.

   C. The difference between the asset with the highest VaR and the asset with the second highest VaR.

Solution:

A correctly defines incremental VaR. Incremental VaR is the change in VaR from increasing a position in an asset, not a change in VaR from an increase in volatility. B is not correct because incremental volatility reflects the results of intentional changes in exposure, not uncontrollable market volatility. C is not correct because incremental VaR is not the difference in the VaRs of the assets with the greatest and second greatest VaRs.

3. Which of the following statements is correct about marginal VaR?

   A. The marginal VaR is the same as the incremental VaR.

   B. The marginal VaR is the VaR required to meet margin calls.

   C. Marginal VaR estimates the change in VaR for a small change in a given portfolio holding.

Solution:

C is correct. In A, marginal VaR is a similar concept to incremental VaR in that they both deal with the effect of changes in VaR, but they are not the same concept. B is incorrect because marginal VaR has nothing to do with margin calls.

7 OTHER KEY RISK MEASURES

- ☐ describe sensitivity risk measures and scenario risk measures and compare these measures to VaR
- ☐ demonstrate how equity, fixed-income, and options exposure measures may be used in measuring and managing market risk and volatility risk
- ☐ describe the use of sensitivity risk measures and scenario risk measures

Just as no single measure of a person's health gives a complete picture of that person's physical condition, no single risk measure gives a full picture of a portfolio's risk profile. As we saw, although VaR has many advantages, it also has many limitations. Therefore, good risk managers will use a comprehensive set of risk tools. In this section, we will look at two additional classes of risk measures: those based on sensitivity analysis and those based on the use of hypothetical or historical scenarios. The former enable us to estimate how our estimated gains and losses change with changes in the underlying risk factors, whereas the latter are based on situations involving considerable market stress from which we estimate how our portfolio will perform.

Sensitivity Risk Measures

Equity, fixed-income, and options positions can be characterized by a number of exposure measures that reflect the sensitivities of these positions to movements in underlying risk factors. Sensitivity measures examine how performance responds to a single change in an underlying risk factor. Understanding and measuring how portfolio positions respond to the underlying sources of risk are primary objectives in managing risk.

Equity Exposure Measures

The primary equity exposure measure is the beta. In a simple world, a single market factor drives equity returns. The return on a stock is given by the familiar capital asset pricing model (CAPM):

$$E(R_i) = R_F + \beta_i[E(R_M) - R_F],$$

where $E(R_i)$ is the expected return on the asset or portfolio i, R_F is the risk-free rate, $E(R_m)$ is the expected return on the market portfolio, and β_i is the beta, which is the risk measure. The expression $E(R_m) - R_F$ is the equity risk premium, which is the return investors demand for investing in equities rather than risk-free instruments. It should be apparent from this often-used equation that beta measures the sensitivity of the security's expected return to the equity risk premium. The beta is defined as the covariance of the asset return with the market return divided by the variance of the market return. The broad market beta, which is an average of all individual betas, is 1.0. Assets with betas more (less) than 1 are considered more (less) volatile than the market as a whole. The CAPM has a number of extensions, including multifactor models, and risk measures derived from those models can also provide more nuanced information on equity risk exposures.

Fixed-Income Exposure Measures

The primary sensitivity exposure measures for fixed-income investments are duration and convexity. (Note that credit, a major factor driving non-government fixed-income markets, is covered elsewhere.) **Duration** is sometimes described as the weighted-average time to maturity of a bond, in which the bond is treated as partially maturing on each coupon payment date. Duration is a sensitivity measure. Under the assumption that all interest rates that affect a bond change by the same percentage, the duration is a measure of the sensitivity of the bond price to the interest rate change that characterizes all rates. This single rate can be viewed as the bond's yield, y. Given a bond priced at B and yield change of Δy, the rate of return or percentage price change for the bond is approximately given as follows:

$$\frac{\Delta B}{B} \approx -D\frac{\Delta y}{1+y},$$

where D is the duration.(The ≈ sign stands for the phrase "approximately equal" and reflects the fact that the relationship is not exact.) In this expression, it is easy to see that duration does reflect the sensitivity of a bond's price to its yield, although under the restrictive assumption of a single change to all rates. The assumption of a single change to all rates may seem fairly restrictive, but ultimately the assumption is encapsulated by assuming that a single discount rate, the yield, drives the bond price. Duration is considered to be a fairly good sensitivity measure. As previously mentioned, duration is a time measure, the weighted-average maturity of a bond, in which the bond is viewed as maturing progressively as it makes its coupon payments.

The relationship shown here is approximate. The formula is derived under the assumption that the yield change is infinitesimally small, and duration fails to accurately capture bond price movements when yield changes are relatively large. Thus, in the above expression, Δy is for small yield changes. It is not possible, however, to say how small a yield change must be before it is small enough for the expression to hold true. In addition, the expression holds only at any instant in time and only for that instant. Over longer periods, the relationship will be less accurate because of the passage of time and because Δy is likely to be larger. To accommodate longer periods of time and larger yield changes, we can incorporate a second factor called **convexity**, which is denoted C. Convexity describes the sensitivity of a bond's duration to changes in interest rates. Adding convexity to the expression, we obtain the following formula:

$$\frac{\Delta B}{B} \approx -D\frac{\Delta y}{1+y} + \frac{1}{2}C\frac{\Delta y^2}{(1+y)^2}.$$

Convexity can play an important role as a risk measure for large yield changes and long holding periods.

Duration and convexity are essential tools in fixed-income risk management. They allow the risk manager to assess the potential losses to a fixed-income portfolio or position under a given change in interest rates.

Options Risk Measures

Derivatives have their own unique exposure measures. Because forwards, futures, and swaps have payoffs that are linear in relation to their underlying, they can often be evaluated using the same exposure measures as their underlying. Options, however, have non-linear payoffs, which result in them having their own family of exposure measures that incorporate this non-linear behavior.

Although options can be very risky instruments in and of themselves, they are a critical tool for effective risk management and are often used to create an exposure to offset an existing risk in the portfolio. The relative riskiness of an option arises from the high degree of leverage embedded in most options. An additional and very important risk can also arise from the sensitivity of an option to the volatility of the underlying security. We will expand on these points in the next few paragraphs.

The most fundamental risk of an option is its sensitivity to the price of the underlying. This sensitivity is called the option's **delta**. Although delta is derived by using mathematics beyond the scope of this reading, we can provide a simple and reasonably effective definition as follows:

$$\Delta \quad (\text{delta}) \approx \frac{\text{Change in value of option}}{\text{Change in value of underlying}}.$$

Call option deltas range from a value of 0 to a value of 1, whereas put option deltas range from a value of 0 to a value of –1. A value of 0 means that the option value does not change when the value of the underlying changes, a condition that is never absolutely true but can be roughly true for a very deep out-of-the-money option. A call delta of 1 means that the price of the call option changes in unison with the underlying, a condition that is also never absolutely true but is *approximately* true for very deep in-the-money calls. A put delta of –1 means that the price of the put option changes in unison with the underlying but in the opposite direction, a condition that is also never absolutely true but is *approximately* true for very deep in-the-money puts. As expiration approaches, an in-the-money call (put) delta approaches 1 (–1) and an out-of-the-money call (put) delta approaches 0.

Delta can be used to approximate the new price of an option as the underlying changes. For a call option, we can use the following formula:

$$c + \Delta c \approx c + \Delta_c \Delta S.$$

Here, c is the original price of the option and Δc is the change in the price. We approximate the change in the price as the product of the call's delta, Δ_c, and the change in the value of the underlying, ΔS. The same relationship would hold for puts, simply changing the c's to p's.

The delta of an option is somewhat analogous to the duration of a fixed-income security. It is a first-order effect, reflecting the direct change in the value of the option or fixed-income security when the underlying price or yield, respectively, changes. Just as duration captures the effect of only small changes in the yield over a short period of time, delta captures the effect of only small changes in the value of the underlying security over a short period of time. Similar to duration, which has the second-order effect of convexity, we can add a second-order effect for options called **gamma**. Gamma is a measure of how sensitive an option's delta is to a change in the underlying. It is a second-order effect in that it is measuring the sensitivity of the first-order effect, delta. Gamma can be interpreted in several ways. The delta reflects the direct change in the value of the underlying position, whereas gamma reflects the indirect change (i.e., the change in the change). Technically, it reflects the change in the delta, as indicated by the following:

$$\Gamma \quad (\text{gamma}) \approx \frac{\text{Change in delta}}{\text{Change in value of underlying}}.$$

As with convexity, gamma itself is not simple to interpret. For example, a call option might have a delta of 0.6 and a gamma of 0.02. It is not easy to determine whether the gamma is large or small. Using the equation just given, if the value of the underlying increases by 0.10 and the gamma is 0.02, then the delta would increase by 0.002 (0.10 × 0.02), from 0.6 to 0.602. Gammas get larger as the option approaches at-the-money, and they are large when options approach expiration, unless the option is deeply in or out of the money. Gamma reflects the uncertainty of whether the option will expire in or out of the money. When an option is close to expiration and roughly at the money, a small change in the price of the underlying will determine whether the option expires worthless or in the money. The uncertainty associated with this win-or-lose situation over a very short time frame leads to a large gamma.

Using delta and gamma, the new call price is

$$c + \Delta c \approx c + \Delta_c \Delta S + \frac{1}{2}\Gamma_c (\Delta S)^2,$$

where Γ_c is the gamma of the call. This equation is similar to the corresponding expression that relates yield changes to bond price changes through duration and convexity. Indeed, as we said, gamma is a second-order effect, like convexity.

A third important sensitivity measure for options is **vega**, and it reflects the effect of volatility. Vega is a first-order effect reflecting the relationship between the option price and the volatility of the underlying. Vega is expressed by the following relationship:

$$\text{Vega} \approx \frac{\text{Change in value of option}}{\text{Change in volatility of underlying}}.$$

Most options are very sensitive to the volatility of the underlying security. The effect of changing volatility can have a material impact on the value of the option, even when the value of the underlying is not changing.

Using delta, gamma, and vega, the new value of an option given an old value, a change in the value of the underlying, and a change in the volatility can be estimated as follows:

$$c + \Delta c \approx c + \Delta_c \Delta S + \frac{1}{2}\Gamma_c (\Delta S)^2 + \text{vega}(\Delta\sigma),$$

where $\Delta\sigma$ is the change in volatility.

The expression represents a composite sensitivity relationship for options. It reflects the expected response of an option value to changes in the value and volatility of the underlying, the two primary factors that change in an unpredictable manner and influence the option value. For portfolios that contain options, understanding these relationships and using them to assess the portfolio's response to market movements are essential elements of effective risk management.

These option measures are applicable not only to options but also to portfolios that contain options. For example, the delta of a portfolio consisting of a long position in an S&P 500 ETF and a short position in a call option on the ETF has a delta that is determined by both the ETF and the option. The ETF has a delta of 1; it changes one-for-one with the S&P 500. The option delta, as noted, has a delta between 0 and 1, though technically 0 and –1 because the option position is short. The ETF has no gamma or vega, so the portfolio gamma and vega are determined by the option. The overall deltas, gammas, and vegas are sums of the deltas, gammas, and vegas of the component positions, taking into account the relative amounts of money invested in each position. Risk managers need to know the overall deltas, gammas, vegas, durations, convexities, and betas to get a comprehensive picture of the sensitivity of the entire portfolio to the prices and volatilities of the underlying.

EXAMPLE 7

Sensitivity Risk Measures

1. Which of the following *most* accurately characterizes duration and convexity?

   A. Sensitivity of bond prices to interest rates

   B. First- and second-order effects of yield changes on bond prices

   C. Weighted-average time to maturity based on the coupon payments and principal

Solution:

B is correct. Duration is the first-order effect and convexity the second-order effect of a change in interest rates on the value of a bond. A and C are correct with respect to duration, but not for convexity.

2. Which of the following statements about the delta of a call option is **not** correct?

   A. It ranges between 0 and 1.

   B. It precisely captures the change in the call value when the underlying changes.

   C. It approaches 1 for an in-the-money option and 0 for an out-of-the-money option.

Solution:

B is correct. A and C correctly characterize delta, whereas B states that delta is precise, which is incorrect because it gives an approximate relationship.

3. Which of the following statements about gamma and vega are correct?

   A. Gamma is a second-order effect, and vega is a first-order effect.

   B. Gamma is the effect of volatility, and vega is the effect of changes in volatility.

   C. Gamma is a second-order effect arising from changes in the sensitivity of volatility to the underlying price.

Solution:

A is correct. B is not correct because gamma does not capture the effect of volatility. Vega is the effect of volatility, but it relates to the level and not the change in volatility. C is incorrect because although gamma is a second-order effect on the option value, it is not related to the sensitivity of volatility to the underlying price.

8 SCENARIO RISK MEASURES

- ☐ describe sensitivity risk measures and scenario risk measures and compare these measures to VaR
- ☐ describe the use of sensitivity risk measures and scenario risk measures

A scenario risk measure estimates the portfolio return that would result from a hypothetical change in markets (a hypothetical scenario) or a repeat of a historical event (a historical scenario). As an example, the risk manager might want to understand how her current portfolio would perform if an event, such as the Black Monday of October 1987, were to reoccur. The factor movements that characterized the historical event would be applied to the factor exposures of the current portfolio. Alternatively, the risk manager may develop a hypothetical scenario to describe a market event that has not occurred in the past but which he or she believes has some probability of occurring in the future. The two elements of scenario risk measures that set them apart from sensitivity risk measures are (1) the use of multiple factor movements used in the scenario measures versus the single factor movements typically used in risk sensitivity measures and (2) the typically larger size of the factor movement used in the scenario measures. Scenario risk measures are related to VaR in that they focus on extreme outcomes, but they are not bound by either recent historical events or assumptions about parameters or probability distributions. **Stress tests**, which apply extreme negative stress to a particular portfolio exposure, are closely related to scenario risk measures. Scenario analysis is an open-ended exercise that could look at positive or negative events, although its most common application is to assess the negative outcomes. Stress tests intentionally focus on extreme negative events to assess the impact of such an event on the portfolio.

The two types of scenario risk measures—historical scenarios and hypothetical scenarios—are discussed in the following sections.

Historical Scenarios

Historical scenarios are scenarios that measure the portfolio return that would result from a repeat of a particular period of financial market history. Historical scenarios used in risk management include such events as the currency crisis of 1997–1998, the market dislocation surrounding the failure of Long-Term Capital Management, the market rout of October 1987, the bursting of the technology bubble in 2001, and the financial crisis of 2008–2009. In order to create a historical scenario, the current set of portfolio holdings is placed into the appropriate valuation models.

Equity positions can often be modeled using their price histories as proxies for their expected behavior, although some practitioners model equities using factor analysis. Valuation models are needed for fixed-income and derivatives products because they have a maturity or an expiration feature that must be accommodated when modeling the portfolio. Historical prices for the fixed-income and derivatives positions currently held in the portfolio may not exist, as in the case of a bond that was issued after the historical period being modeled. Even when historical prices for specific instruments do exist, they may not be relevant to the current characteristics of the instrument. Take the case of a 5-year historical price series for a 10-year bond with 1 year remaining to maturity; the historical price series reflects the price volatility of what used to be a longer bond (e.g., five years ago, the bond had six years remaining to maturity; three years ago, the bond had four years remaining to maturity). The volatility of the bond when it had six years remaining to maturity would be higher than it is today, with only one year remaining to maturity. Using its historical price history would mischaracterize the risk of the current portfolio holding. For this reason, the historical yields, spreads, implied volatilities, prices of the underlying assets in derivatives contracts, and the other input parameters that drive the pricing of these instruments are more important in explaining the risks of these instruments than the price history of the instrument itself.

Some examples may help to show how fixed-income or derivatives valuation models are used in a historical scenario. In the case of a convertible bond, the bond's terms and conditions (e.g., coupon, conversion ratio, maturity) are entered into a convertible bond pricing model. In the case of standard bonds, the terms and conditions of these instruments (e.g., coupon, call features, put features, any amortization or sinking fund features, maturity) are entered into fixed-income pricing models. These modeled fixed-income or derivatives holdings, together with the equity holdings, are then re-priced under the conditions that prevailed during the "scenario period"—a given set of dates in the past. Changes in interest rates, credit spreads, implied volatility levels, and any asset underlying a derivatives product, as well as the historical price changes in the equity portfolio, would all be reflected in the re-priced portfolio. The value of each position is recorded before and after these changes in order to arrive at the gain or loss that would occur under the chosen scenario. Historical scenario events are specifically chosen to represent extreme market dislocations and often exhibit abnormally high correlations among asset classes. It is most common to run the scenario or stress test as if the total price action movement across the period occurs instantaneously, before any rebalancing or management action is possible. The output of the scenario can include

- the total return of the portfolio;
- for long-only asset managers, the total return of the portfolio relative to its benchmark;
- for pensions, insurers, and others whose liabilities are not already incorporated into the portfolio, the total return of the portfolio relative to the change in liabilities under the scenario; and
- any collateral requirements and other cash needs that will be driven by the changes specified in the scenario.

One variation of the historical scenario approach includes running the scenario over multiple days and incorporating actions that the manager might be expected to take during the period. Instead of assuming the shock is a single instant event, this approach assumes it takes place over a number of days and that on each day the portfolio manager can take such actions as selling assets or rebalancing hedges.

Many risk managers are skeptical of this approach because it produces smaller potential loss measures (by design) and does not answer important questions that have been relevant in real crises, such as, "What if the severe price action happens so quickly that the portfolio manager cannot take remedial actions?" Generally, risk managers prefer that a stress testing exercise be tailored to the *initial outcome of a large shock*, to ensure that the event is survivable by a portfolio that uses leverage, and that there will be no unacceptable counterparty exposures or portfolio concentrations before action can be taken to improve the situation. This method also helps to simulate the possibility that liquidity may be unavailable.

Risk managers seeking to measure the impact of a historical scenario need to ensure all relevant risk factors are included. For instance, foreign equities will need to be decomposed into foreign exchange exposure and equity exposure in the analysis. Stress tests typically take the explicit currency approach, which measures the currency exposure of each foreign equity. Alternatively, the risk manager may use an approach that incorporates implicit currency risks, such as companies that may be registered in one country but have earnings flowing in from other countries, and may hedge some of those revenues back to their base currency.

When the historical simulation fully revalues securities under rate and price changes that occurred during the scenario period, the results should be highly accurate. Sometimes, however, scenarios are applied to risk sensitivities rather than the securities themselves. This approach is a simpler form of analysis, but it should not be used for options or option-embedded securities. Although it may be tempting to use delta and gamma or duration and convexity to estimate the impact of a scenario on options or option-embedded securities, these measures are not suited for handling the kinds of extreme movements analyzed in scenario analysis. Although gamma and convexity are second-order adjustments that work with delta and duration to estimate extreme movements, they are inadequate for scenario analysis.

Even in simpler fixed-income cases in which no options are present, care needs to be taken to ensure the analysis does not oversimplify. Duration sensitivities can be used as the inputs to a scenario analysis for straightforward fixed-income instruments, but these sensitivities need to be mapped to the most relevant sectors, credit curves, and yield curve segments before beginning the analysis. If assets are mapped too broadly, the analysis will miss the important differences that could drive the most meaningful outcomes in a given scenario.

It is also important to pay careful attention to how securities or markets that did not yet exist at the time of the scenario are modeled. If, for instance, an analyst is measuring a current portfolio's sensitivity to a recurrence of the 1987 US stock market crash, the analyst needs to determine how to treat stocks in the portfolio that had an initial public offering after 1987. They may need to be mapped to a relevant index or to a similar company or be decomposed into the relevant statistical factors (such as growth, value, volatility, or momentum) by using a factor model before beginning the analysis. Similarly, because credit default swaps did not come into widespread use until 2002, historical scenarios for dates preceding this time would need to be adapted to appropriately reflect the impact of a repeat of that scenario on these new securities.

Hypothetical Scenarios

Scenarios have a number of benefits. They can reflect the impact of extreme market movements, and they make no specific assumptions regarding normality or correlation. Historical scenarios have the extra benefit of being uncontroversial; no one can claim it is impossible for such events to occur, because they did. One problem with scenario analysis, however, lies in ascribing the probability of a given scenario. Most would agree that it is improbable to assume that the exact historical scenario specified will actually occur in precisely the same way in the future. Another potential problem is that, because it has happened (particularly when it has happened recently), risk managers or portfolio managers are inclined to take precautions that make their portfolios safer for a replay of that historical crisis—and, in the process, make their portfolios more vulnerable to a crisis that has not yet happened.

For that reason, risk managers also use hypothetical scenarios—extreme movements and co-movements in different markets that have not necessarily previously occurred. The scenarios used are somewhat difficult to believe, and it is difficult to assess their probability. Still, they represent the only real method to assess portfolio outcomes under market movements that might be imagined but that have not yet been experienced.

To design an effective hypothetical scenario, it is necessary to identify the portfolio's most significant exposures. Targeting these material exposures and assessing their behavior in various environments is a process called **reverse stress testing**. The risk manager is seeking answers to such questions as the following: What are the top 10 exposures or risk drivers in my portfolio? What would make them risky? What are the top 10 benchmark-relative exposures? Under what scenario would hedges not hedge? Under what scenario would my securities lending activity, ordinarily thought to be riskless, be risky? The ideal use of hypothetical scenarios is, then, not to model every possible future state of every market variable, but rather to target those that are highly significant to the portfolio in order to assess, and potentially address, vulnerabilities.

Reverse stress testing is particularly helpful in estimating potential losses if more than one important exposure is affected in a market crisis, as often happens when participants "crowd" into the same exposures. Sometimes, apparently unrelated markets experience stress at the same time.

The risk manager might also choose to design a hypothetical geopolitical event, estimating its potential effect on markets and the resulting impact on the portfolio. To develop these scenarios, individuals with varying areas of expertise posit an event—such as an earthquake in Country Y, or Country X invades Country Z, or the banking system implodes in Region A. The group conducting the analysis identifies which markets are most likely to be affected as well as any identifiable secondary effects. The next step is to establish a potential range of movement for the affected markets.

The final scenario is intended to meet the standard of "rare, but not impossible." The exercise is unlikely to be truly accurate in the face of the real event, but it will often help to identify unexpected portfolio vulnerabilities and outcomes and to think through counterparty credit and operational considerations that could exacerbate or accelerate the scenario.

Hypothetical scenarios are particularly beneficial in being able to stress correlation parameters. The scenario is not constrained to assume that assets will co-move as they have done in the past, which can help identify dangers that other forms of risk analysis may miss. Scenarios can be designed to highlight that correlations often increase in times of stress. This is often achieved by subjecting markets that typically have little or no correlation with one another to the same or similar movements, thereby simulating a temporarily higher correlation. Scenarios can also be devised to pinpoint times when hedging might work poorly—when assets, such as a bond and the credit default swap used to hedge it, that normally have a high correlation might temporarily decouple and move by different percentages or even in different directions. This often occurs when markets experience a "flight to quality"; the swap rate may move down as a result of their relative credit strength, whereas the bond yield might increase given its perceived credit risk.

Once a risk manager has completed a scenario analysis, common questions may be, "What do you do with a scenario analysis? What are the action steps?" If the portfolios are within all other rules and guidelines—their exposures have been kept within desired limits and their VaR or *ex ante* tracking error is within the desired range—scenario analysis provides one final opportunity to assess the potential for negative surprises during a given stress event. The action steps might be to trim back positions that are otherwise within all limits and that appear to present comfortable risk exposures under the current environment but would perform unacceptably during a plausible stress environment. In the case of asset management, where clients have elected to be in a given asset class and the asset manager is constrained by that investment mandate, action steps may include adjusting benchmark-relative risk, disclosing to clients the manager's concerns regarding the risks in the portfolio, or changing counterparty or operational procedures to avoid an unwanted event.

But a caution is in order: A portfolio that has no sensitivity to any stress event is unlikely to earn more than the risk-free rate, or in the case of long-only asset managers, outperform the benchmark index. Stress tests and scenarios analyses are best used in the effort to *understand* a portfolio's risk exposures, not to eliminate them. Effective risk management sets a tolerance range for a stress test or scenario that reflects a higher loss possibility than the investment manager would normally find acceptable. Scenarios should be periodically run again, and action should be taken only if the portfolio exceeds this relatively high tolerance level. It is also important to continually evaluate new threats and new market developments and to periodically refresh the set of scenarios, removing scenarios that are no longer meaningful for the portfolio.

Note also that scenario risk measures and stress tests are best used as the final screen in a series of position constraints that include position size limits, exposure limits, and VaR or *ex ante* tracking error limits. They do not serve well as the initial or primary screen, for reasons that will be discussed shortly.

Parties that use leverage, such as banks and hedge funds, are more likely to use single-factor stress tests rather than multifactor scenario analyses. The focus on a single factor helps in assessing whether a given exposure is likely to impair their capital under a given stress movement; these are pass/fail tests. If capital falls below an acceptable level, it could set off a chain reaction of margin calls, withdrawal of financing, and other actions that threaten the viability of the business.

EXAMPLE 8

Scenario Analysis

1. Which of the following is an example of a reverse stress test?

   A. Identify the top 10 exposures in the portfolio, and then generate a hypothetical stress that could adversely affect all 10 simultaneously.

   B. Find the worst single day's performance that could have occurred for the current portfolio had it been held throughout the past five years.

   C. Find the returns that occurred in all risk factors in the 2008 global financial crisis, reverse the sign on these, and apply them to today's portfolio.

Solution:

A is correct. B is not a reverse stress test because reverse stress tests focus more narrowly on trouble spots for a specific portfolio. C would illustrate how the portfolio would have performed in an extremely strong market, quite unlike what occurred in 2008.

2. Which kind of market participant is *least likely* to use scenario analysis as a pass/fail stress test?

   A. Bank

   B. Long-only asset manager

   C. Hedge fund using leverage

Solution:

B is correct. Long-only asset managers do not typically use leverage and are thus less likely to become insolvent, making a pass/fail test for solvency less relevant to them. A and C are not correct because parties that use leverage, such as hedge funds and banks, are likely to use stress tests to determine what market movements could impair their capital and lead to insolvency.

3. What is the *most* accurate approach to scenario analysis for a portfolio that uses options?

   A. Apply the scenario to option delta.

   B. Apply the scenario to option delta + gamma.

   C. Fully reprice the options using the market returns specified under the scenario.

Solution:

C is correct. Both A and B risk misestimating the actual results of the scenario because both delta and gamma estimate how an option's value might change for a small move in the underlying asset, not the large movements typically used in a scenario analysis.

9 SENSITIVITY AND SCENARIO RISK MEASURES AND VAR

- ☐ describe sensitivity risk measures and scenario risk measures and compare these measures to VaR
- ☐ describe advantages and limitations of sensitivity risk measures and scenario risk measures

Although both VaR and sensitivity risk measures deal with related concepts, they have their own distinctions. VaR is a measure of losses and the probability of large losses. Sensitivity risk measures capture changes in the value of an asset in response to a change in something else, such as a market index, an interest rate, or an exchange rate; they do not, however, tell us anything about the probability of a given change in value occurring. For example, we could use duration to measure the change in a bond price for an instantaneous 1 bp change in the yield, but duration does not tell us anything about the likelihood of such a change occurring. Similar statements could be made about equities and the various option measures: Betas and deltas do not tell us how likely a change might be in the underlying risk factors, but given a change, they tell us how responsive the asset or derivative would be.

VaR gives us a broader picture of the risk in the sense that it accounts for the probability of losses of certain amounts. In this sense, it incorporates what we know about the probability of movements in the risk factors. Nonetheless, these sensitivity measures are still very useful in that they allow us to take a much more detailed look at the relationships driving the risk. It is one thing to say that a VaR is $2 million for one day at 5%. We know what that means. But it is equally important to understand what is driving the risk. Is it coming from high beta stocks, high duration bonds, or high delta options? If we find our VaR unacceptable, we have to know where to look to modify it. If we simply use VaR by itself, we will blindly rely on a single number without understanding what factors are driving the number.

VaR has much in common with scenario risk measures in that both types of measures estimate potential loss. VaR tends to do so using a model for which input parameters are created based on market returns from a particular time in history. Thus, the VaR estimate is vulnerable if correlation relationships and market volatility during the period in question are not representative of the conditions the portfolio may face in the future. VaR does, however, allow a disciplined method for stressing all factors in the portfolio. Scenario analysis allows either the risk assessment to be fully hypothetical or to be linked to a different and more extreme period of history, helping reduce some of the biases imposed by the VaR model. But there is no guarantee that the scenario chosen will be the "right" one to estimate risk for future markets. Moreover, it is particularly difficult to stress all possible risk factors in a hypothetical scenario in a way that does not embed biases similar to those that occur in VaR modeling.

Each of these measures—sensitivity risk measures, scenario risk measures, and VaR—has distinct limitations and distinct benefits. They are best used in combination because no one measure has the answer, but all provide valuable information that can help risk managers understand the portfolio and avoid unwanted outcomes and surprises.

Advantages and Limitations of Sensitivity Risk Measures and Scenario Risk Measures

Before portfolios began using risk measures based on modern portfolio theory, the very first risk measure was "position size"—the value invested in a given type of asset. Position size is a very effective risk measure for homogeneous, long-only portfolios, particularly for those familiar with the homogenous asset class in question; an experienced person can assess what the loss potential of such a portfolio is just by knowing its size. But position size is less useful for assessing interest rate risk, even less useful for summarizing the risk of a multi-asset class portfolio, and less useful still at assessing net risk in a portfolio that uses hedging instruments, short positions, and liabilities.

Sensitivity measures address some of the shortcomings of position size measures. Duration, for example, addresses the difference between a 1-year note and a 30-year note; it measures the level of interest rate risk. Option delta and duration (for fixed income) help to display net risk in a portfolio that has hedging or short positions with optionality or interest rate risk.

Sensitivities typically do not often distinguish assets by volatility, though. When measured as the sensitivity to a 1 bp or 1% move, they do not tell the user which portfolio has greater loss potential any more than position size measures do. A high-yield bond portfolio might have the same sensitivity to a 0.01% credit spread movement as an investment-grade portfolio, but they do not have the same risk because the credit spreads of the high-yield portfolio are more likely to move 0.01%, or more, than the credit spreads of the investment-grade bonds. Sensitivity measures do not distinguish by standard deviation/volatility or other higher confidence loss measures. Measuring sensitivity to a one standard deviation movement in an asset's price or yield, however, is one way to overcome this shortcoming of sensitivity.

Granularity: Too Much or Too Little?

Sensitivity measures are aggregated in categories or "buckets." (A bucket is a risk factor description such as "one- to five-year French sovereign debt.") When a number of fixed-income positions are assigned to the same bucket, the effect is an assumption of perfect correlation across the risks encompassed by that bucket. For the "one- to five-year French sovereign debt" risk factor, a short duration position in four-year French sovereign debt will be assumed to fully offset a long duration position in two-year French sovereign debt. However, this may not be true in the case of a non-parallel interest rate change; these points on the yield curve do not have a correlation coefficient of 1 to one another. The broader the buckets used, the more they can hide this kind of correlation risk; but the narrower the buckets used, the greater the complexity and thus the more difficult to portray portfolios in simple, accessible ways. The width or the narrowness of the risk-factor buckets used to portray sensitivity measures is referred to as granularity.

Scenario analysis and stress testing have well-deserved popularity, and they address many of the shortcomings of VaR described earlier. Sensitivity and scenario risk measures can complement VaR in the following ways:

- They do not need to rely on history. Sensitivity and scenario risk measures can be constructed to test the portfolio's vulnerability to a truly never-before-seen market movement. In this way, they can be free of the volatility and correlation behavior of recent market history, which may simply not be representative of stress conditions. In a scenario analysis, assets that typically have a low correlation with one another can be modeled under an assumption of perfect positive correlation simply by simulating an identical price movement for these assets. Alternatively, they can be modeled under an assumption of perfect negative correlation by simulating identical price movements (i.e., in the opposite direction). A scenario might be designed in which a market that typically exhibits an annual standard deviation of 15% moves by 20% in a single day.

- Scenarios can be designed to overcome any assumption of normal distributions; the shock used could be the equivalent of 1, 10, or 1,000 standard deviations, at the choice of the analyst—or as provided by an actual moment in history.
- Scenarios can be tailored to expose a portfolio's most concentrated positions to even worse movement than its other exposures, allowing liquidity to be taken into account.

But scenario measures are not without their own limitations:

- Historical scenarios are interesting, and illuminating, but are not going to happen in exactly the same way again, making hypothetical scenarios necessary to truly fill the gaps identified with the other risk measures listed.
- Hypothetical scenarios may incorrectly specify how assets will co-move, they may get the magnitude of movements wrong, and they may incorrectly adjust for the effects of liquidity and concentration.
- Hypothetical scenarios can be very difficult to create and maintain. Getting all factors and their relationships accurately represented in the suite of scenarios is a painstaking and possibly never-ending exercise. Accordingly, it is necessary to draw a line of "reasonableness" at which to curtail the scenario analysis, and by the very act of being curtailed, the scenario might miss the real risk.
- It is very difficult to know how to establish the appropriate limits on a scenario analysis or stress test. Because we are proposing hypothetical movements in markets and risk factors, we cannot use history to assign a probability of such a move occurring. What if rates rise instantaneously 0.50%, 1.00%, or 3.00%? How should the short end of the yield curve move versus the long end? How much should credit spreads of different qualities move? It is difficult to choose.

The more extreme the scenario, and the farther from historical experience, the less likely it is to be found believable or actionable by management of a company or a portfolio. This issue tends to lead scenario constructors to underestimate movement in order to appear credible. As an example, prior to the very large drop in real estate values that prevailed in the United States from 2008 to 2010, no similar nationwide price decline had occurred in history. Risk measurement teams at a number of firms did prepare scenarios that estimated the potential outcome if real estate prices declined meaningfully, but their scenarios in many cases were only half as large as the movements that subsequently occurred. Because these large market movements had never before occurred, there was no historical basis for estimating them, and to do so appeared irresponsible. This is an additional risk of scenario analysis: The need to keep the scenario plausible may lead to it being incorrect.

In sum, scenario analyses and stress tests have the opportunity to correct the failings of probabilistic risk measures, such as VaR and *ex ante* tracking error; however, because the version of the future they suggest may be no more accurate than that used in VaR, they may also fail to predict potential loss accurately.

As we can see, each risk measure has elements that are better than the others, and each has important failings. No one measure is the "solution" to risk management. Each is useful and necessary to answer certain questions but not sufficient to answer all possible questions—or to prevent all forms of unexpected loss. Using the measures in combination, to correct each other's failings, is as close to a solution as we come. Designing constraints by using multiple measures is the key practice used by successful risk managers. Viewing a portfolio through these multiple lenses provides a more solid framework for a risk manager or an investor to exercise judgment and can help reduce conceptual bias in portfolio management.

EXAMPLE 9

Limitations of Risk Measures

1. Which of the following is **not** a limitation of VaR?

   A. It does not adjust for bonds of different durations.

   B. It largely relies on recent historical correlations and volatilities.

   C. It can be inaccurate if the size of positions held is large relative to available liquidity.

Solution:

A is correct. Well-executed VaR measures do adjust for bonds of differing duration, and therefore it is not a limitation of VaR. B is incorrect because VaR ordinarily uses some period of recent history as part of the calculation, and this reliance on history is one of its limitations. C is incorrect because VaR can be inaccurate and underestimate risk if portfolio positions are too large relative to the available market liquidity, and this inability to account for the illiquidity of an individual investor's position is an additional limitation of VaR.

2. Which of the following statements about sensitivities is true?

   A. When duration is measured as the sensitivity to a 1 bp change in interest rates, it can be biased by choice of the historical period preceding this measure.

   B. Sensitivity measures are the best way to determine how an option can behave under extreme market movements.

   C. Duration effectively assumes that the correlation between a fixed-income exposure and the risk-free rate is 1, whereas beta takes into account the historical correlation between an equity and its comparison index.

Solution:

C is correct. Duration assumes that all interest rates that affect a bond change by the same percentage (an effective correlation of 1). A is incorrect because the 1 bp change in rates is applied to current rates, not historical rates. B is incorrect because sensitivity measures are often too small to reveal the most extreme movements for option positions; the larger shocks used in scenario measures are preferable to reveal option characteristics.

3. Which of the following is **not** a limitation of scenario measures?

   A. It is difficult to ascribe probability to a given scenario.

B. Scenario measures assume a normal distribution, and market returns are not necessarily normal.

C. They risk being an infinite task; one cannot possibly measure all of the possible future scenarios.

Solution:

B is correct. Scenario measures do not assume any given distribution, and thus this is not a limitation of scenario analysis. A is incorrect because it is in fact difficult to ascribe probability to many scenarios, and thus this is a limitation of scenario analysis. C is also incorrect because it is in fact impossible to measure all possible future scenarios, and this is a limitation of scenario analysis.

4. Which measures are based on market returns during a particular historical period?

A. Hypothetical scenario analysis and duration sensitivity

B. Historical scenario analysis and VaR

C. Option delta and vega

Solution:

B is correct. Historical scenarios apply market returns from a particular period to the portfolio, and virtually all VaR methodologies use a historical period to underpin the VaR model (although certain methods may make adjustments if this historical period is seen to be anomalous in some way). A is incorrect because a hypothetical scenario is not based on an actual historical period, and duration sensitivity measures change in value for a given small change in rates, not for a given historical period. C is incorrect because option delta and vega measure how much an option's value will change for a given change in the price of the underlying (delta) or implied volatility (vega), and these are sensitivity measures, not measures based on a particular historical period.

10 USING CONSTRAINTS IN MARKET RISK MANAGEMENT

- ☐ explain constraints used in managing market risks, including risk budgeting, position limits, scenario limits, and stop-loss limits
- ☐ explain how risk measures may be used in capital allocation decisions

Designing suitable constraints to be used in market risk management is essential to managing risk effectively. Risk *measurements* in and of themselves cannot be said to be restrictive or unrestrictive: The *limits* placed on the measures drive action. VaR can be measured to a very high confidence level (for example, 99%) or to a low level (for example, 84%). But placing a loose limit on a 99% confidence VaR measure could be less of a constraint than placing a tight limit on an 84% confidence measure. It is not the confidence interval that drives conservatism as much as the limit that is placed on it.

If constraints are too tight, they may limit the pursuit of perceived opportunities and shrink returns or profitability to a sub-optimal level. If constraints are too loose, outsized losses can occur, threatening the viability of the portfolio or business. The concept of "restrictive" or "unrestrictive" relates to the risk appetite of the firm or portfolio and the sizes of losses it can tolerate. Unrestrictive limits are typically set far from current risk levels and permit larger losses than restrictive limits. As an example, for a leveraged portfolio in which insolvency could occur if cumulative daily losses exceed $10 million and the portfolio's current two week, 1% VaR measure is $3 million, an unrestrictive limit might be one set at $10 million. If the portfolio increased positions and went right up to its limit, a misestimation of VaR could result in insolvency; moreover, the fact that losses are expected to exceed the measure at least 1% of the time could mean disaster. But if the limit were set at $4 million, the portfolio might under-allocate the capital it has to invest and fail to make a high enough return on equity to thrive in a competitive environment.

Before applying constraints, particularly those involving such potential loss measures as VaR or a scenario analysis, it is worth considering how far down in the organizational hierarchy to impose them. If applied exclusively to lower level business units, the firm's aggregate risk exposure fails to take advantage of offsetting risks that may occur at higher levels of the organization. As a result, the overall company may never be able to invest according to its risk tolerance because it is "stopped out" by rules lower in the organization. For example, imagine a bank with five trading desks: It might have an overall VaR tolerance of €10 million and might set each trading desk's limit for its standalone VaR at €2 million, which seems reasonable. If there is anything lower than perfect correlation across these desks' positions, however—and particularly if one desk has a short position that to some degree serves as an offset to another desk's long position—the firm will never be able to use its €10 million risk appetite in full. The cure for this problem is over-allocation, with the caveat that a given desk might need to be cut back to its pro rata share in the event that correlations among trading desks are higher than, or the short positions across the different portfolios are not as offsetting as, the over-allocation assumes. Alternatively, some firms might use marginal VaR for each trading desk, allocating each desk a VaR budget such that the total VaR is the sum of each individual desk's marginal VaR. This approach permits each trading desk to "reinvest" the diversification benefits obtained at the aggregate level.

Among the constraints most often used in risk management are risk budgeting, position limits, scenario limits, and stop-loss limits. As is the case in risk measurement, for which multiple measures work better than any one measure alone does, so it is in risk constraints. No one approach on its own works perfectly; they are most effective in combination.

Risk Budgeting

In **risk budgeting**, the total risk appetite of the firm or portfolio is agreed on at the highest level of the entity and then allocated to sub-activities. Risk budgeting typically rests on a foundation of VaR or *ex ante* tracking error.

A bank might establish a limit on total economic capital or VaR and describe this limit as its risk appetite. Next, it might allocate this risk appetite among the basic risk types (market, credit, and operational) and different business units, geographies, and activities. It allocates to the business unit and/or risk type by specifying a limit, using its chosen measure, for that given activity. For example, it might allow its European business to use 20% of its market risk capital (the portion of its economic capital expected to be used to support market risk taking) and 40% of its credit risk capital, whereas its Asian business might have a different limit. It will set these limits based on the expected long-term profitability of the opportunity set and the demonstrated skill of a business at delivering profitable results, taking into consideration shareholders' expectations regarding the activities the bank is engaged in. As an example of potential shareholder expectations, consider a case in which a firm's shareholder disclosure suggests that the firm's predominant market risk-taking activities are in the Asian markets and that less risk-taking activity is in Europe. Shareholders will be surprised if greater losses are incurred from its European business than its Asian business. Market risk capital limits for the European business should be lower than for the Asian business to be consistent with shareholder disclosures.

A pension fund sponsor might begin with its tolerance for how much of a mismatch it is willing to tolerate overall between the total value of assets and its liabilities—its surplus at risk. Surplus at risk can be the starting point for its asset allocation decision making. Once the broad asset allocation is established, usually expressed via a set of benchmarks, the pension fund sponsor might further establish its tolerance for underperformance in a given asset class and allocate that tolerance to the asset managers selected to manage the assets by assigning each an *ex ante* tracking error budget.

A portfolio manager might have an *ex ante* tracking error budget explicitly provided by the client, or if none is provided by the client, it might instead develop a tracking error budget based on her investment philosophy and market practice. Given this budget, she will seek to optimize the portfolio's exposures relative to the benchmark to ensure that the strategies that generate the most tracking error for the portfolio are those for which she expects the greatest reward.

Position Limits

Risk budgeting follows a clear logic; but as we have noted, VaR-based measures have a number of drawbacks. One of them is that they perform poorly if portfolios are unusually concentrated, particularly with respect to market liquidity.

Position limits are limits on the market value of any given investment, or the notional principal amount for a derivatives contract. They can be expressed in currency units or as a percentage of some other value, such as net assets. Position limits do not take into account duration, volatility, and correlation, as VaR does, but they are excellent controls on overconcentration. Like risk budgeting, position limits need to be used carefully; if every asset type that a portfolio manager could invest in is constrained, he will have no room to succeed in outperforming the benchmark or generating absolute returns, assuming that is the mandate. Position limits should not be overly prescriptive but should address the event risk and single name risk that VaR handles so poorly, such as

- limits per issuer;
- limits per currency or country;
- limits on categories expected to be minimized in a given strategy, such as high-yield credit or emerging market equities;
- limits on gross size of long–short positions or derivatives activity; and
- limits on asset ownership that correspond to market liquidity measures, such as daily average trading volume.

Scenario Limits

A scenario limit is a limit on the estimated loss for a given scenario, which if exceeded, would require corrective action in the portfolio.

As discussed in Section 3.3, scenarios also address shortcomings of VaR, such as the potential for changes in correlation or for extreme movements that might not be predicted using a normal distribution or the historical lookback period used for the VaR measure. Just producing scenario analysis, however, without having any related action steps is not a very valuable exercise.

The action steps that generally follow a scenario analysis are to examine (1) whether the results are within risk tolerance and, in the case of asset managers, (2) whether the results are well incorporated into investor disclosures. To determine whether results are within the established risk tolerance, a tolerance level for each scenario must be developed. It is better to establish a higher tolerance for potential loss under the most extreme scenarios. If the same limit is applied to all scenarios, even extremely unlikely scenarios (e.g., "interest rates rise 1,000,000%"), then the portfolio will simply not be able to take any risk. The risk manager then observes over time whether the portfolio's sensitivity to the scenario is increasing or crosses this high-tolerance bound.

Stop-Loss Limits

A **stop-loss limit** requires a reduction in the size of a portfolio, or its complete liquidation, when a loss of a particular size occurs in a specified period.

One of the limitations of VaR described in Section 2.3.2 was "trending," in which a portfolio remains under its VaR limit each day but cumulatively loses more than expected. This trending can be managed by imposing and monitoring stop-loss limits in addition to the VaR constraints. In one form of a stop-loss limit, the portfolio's positions are unwound if its losses over a pre-specified period exceed a pre-specified level. (Those levels are typically defined to align with the overall risk tolerance.) As an example, a portfolio might have a 10-day, 1% VaR limit of $5 million, but it will be liquidated if its cumulative monthly loss ever exceeds $8 million. The relationship between the stop-loss and the VaR measure can vary depending on management preferences as well as the differing time periods with which the measures are specified.

An alternative approach to a stop-loss limit might instead be to impose a requirement to undertake hedging activity, which may include purchases of protective options, after losses of a given magnitude, with the magnitude of the hedge increasing as losses increase. This approach, called drawdown control or portfolio insurance, is more dynamic and more sophisticated than the simpler stop-loss limit.

Risk Measures and Capital Allocation

In market risk management, capital allocation is the practice of placing limits on each of a company's activities in order to ensure that the areas in which it expects the greatest reward and has the greatest expertise are given the resources needed to accomplish their goals. Allocating capital wisely ensures that an unproven strategy does not use up all of the firm's risk appetite and, in so doing, deprive the areas most likely to be successful of the capital they need to execute on their strategy.

Economic capital is often used to estimate how much of shareholders' equity could be lost by the portfolio under very unfavorable circumstances. Capital allocation may start with a measurement of economic capital (the amount of capital a firm needs to hold if it is to survive severe losses from the risks in its businesses). The company's actual, physical on-balance-sheet capital must exceed the measure of economic capital, and a minimum level of economic capital must be established to ensure that the company does not take on a risk of loss that will exceed its available capital. The company first establishes its overall risk appetite in economic capital terms, and then it subdivides this appetite among its units. This exercise is similar to risk budgeting, but in the case of corporations, banks, insurers, or hedge funds, it is more likely to be called "capital allocation." Capital allocation is often used in cases in which leverage is used by the portfolio or in which the strategy has meaningful **tail risk**, meaning that losses in extreme events could be far greater than would be expected for a portfolio of assets with a normal distribution. Economic capital is designed to measure how much shareholders' equity could be required to meet tail risk losses. Strategies that have greater-than-expected tail risk include those that sell options, sell insurance, take substantial credit risk, or have unique liquidity or exposure concentration risks. Although risk budgeting more commonly focuses on losses at the one standard deviation level, capital allocation focuses on losses at a very high confidence level in order to capture the magnitude of capital that is placed at risk by the strategy. Capital allocation seeks to understand how much of an investor's scarce resources are, or could be, used by a given portfolio, thereby making it unavailable to other portfolios.

Because a company's capital is a scarce resource and relatively expensive, it should be deployed in activities that have the best chance of earning a superior rate of return. It also should be deployed in a way that investors expect, in activities in which the company has expertise, and in strategies that investors believe the company can successfully execute.

To optimize the use of capital, the "owner" of the capital will typically establish a hurdle rate over a given time horizon; this is often expressed as the expected rate of return per unit of capital allocated. Two potential activities, Portfolio A and Portfolio B, might require different amounts of capital. Portfolio A might require €325,000, and its expected return might be €50,000 per year (15.4%). Portfolio B might have a reasonable expectation of earning €100,000 per year, but it might require €1,000,000 in capital (a 10% return). If the investor has an annualized hurdle rate of 15%, Portfolio A will exceed the hurdle rate and appear a better user of capital than Portfolio B, even though the absolute income for Portfolio B is higher.

Beyond measuring and limiting economic capital, capital allocation is sometimes used as a broad term for allocating costly resources. In some cases, the costly resource is cash; if, for instance, the portfolio has invested in options and futures trading strategies that require heavy use of margin and overcollateralization, its use of economic capital could be low and available cash may be the constraining factor. For other types of investors, such as banks or insurance companies, the capital required by regulatory bodies could be relatively large; as a result, these capital measures may be the most onerous constraint and thus the basis of capital allocation.

When the current measure of economic capital is a smaller number than the portfolio's cash or regulatory capital needs, it may not be the binding constraint. But when it is higher than other measures, it can become the binding constraint, and the one to which hurdle rates should be applied.

EXAMPLE 10

Creating Constraints with Risk Measures

1. Which of the following is **not** an example of risk budgeting?

   A. Giving a foreign exchange trading desk a VaR limit of $10 million

   B. Allowing a portfolio manager to have an *ex ante* tracking error up to 5% in a given portfolio

   C. Reducing the positions in a portfolio after a loss of a 5% of capital has occurred in a single month

Solution:

C is correct. This is an example of a stop-loss limit, not risk budgeting. The other choices are both examples of risk budgeting.

2. Which statement is true regarding risk budgeting in cases in which marginal VaR is used?

   A. The total risk budget is never equal to the sum of the individual sub-portfolios' risk budgets.

   B. The total risk budget is always equal to the sum of the individual sub-portfolios' risk budgets.

   C. If the total risk budget is equal to the sum of the individual sub-portfolios' risk budgets, there is a risk that this approach may cause capital to be underutilized.

Solution:

B is correct. When using marginal VaR, the total risk budget will be equal to the sum of the individual risk budgets. Choice A is not correct. C is also incorrect; it would be correct if each sub-portfolio's individual VaR measure, not adjusted for its marginal contribution, were used, which could lead to underutilization of capital.

MARKET PARTICIPANTS AND THE RISK MEASURES THEY USE

11

- [ ] describe risk measures used by banks, asset managers, pension funds, and insurers

In this section, we examine the practical applications of risk measures. First, we will look at how different types of market participants use risk measures. An understanding of how various market participants use these measures will help as we move to a discussion of their limitations.

Market Participants and the Different Risk Measures They Use

Three factors tend to greatly influence the types of risk measures used by different market participants:

- The degree to which the market participant is leveraged and the resulting need to assess minimum capitalization/maximum leverage ratios;
- The mix of risk factors to which their business is exposed (e.g., the degree of equity or fixed-income concentration in their portfolios);
- The accounting or regulatory requirements that govern their reporting.

Market participants who use a high degree of leverage typically need to assess their sensitivity to shocks to ensure that they will remain a going concern under very severe, but foreseeable, stresses. This leads them to focus on potential loss measures with a high confidence interval or to focus on rare events that might occur in a short period of time, such as two weeks. Those who use minimal (or no) leverage, such as long-only asset managers, are interested in shock sensitivity as well, but they are likely less concerned with trying to discern the difference between a 99.99% (0.01% VaR) worst case and a 99.95% (0.05% VaR) worst case. Their focus is more likely on avoiding underperformance—for example, failing to keep pace with their market benchmark when markets are doing well. For this reason, they are often more interested in lower confidence intervals—events that are more likely to occur and lead to underperformance for a given strategy. Unleveraged asset managers may also prefer to measure potential underperformance over longer periods of time, such as a quarter or a year, rather than shorter periods.

For portfolios dominated by fixed-income investments, risk managers focus on how sensitive the portfolios are to instantaneous price and yield changes in a variety of categories and typically emphasize duration, credit spread duration, and key rate duration measures. Credit spread duration measures the impact on an instrument's value if credit spreads move while risk-free rates remain unchanged. Key rate duration (sometimes called partial duration) measures the sensitivity of a bond's price to changes in specific maturities on the benchmark yield curve. Risk measurement for fixed-income portfolios is conducted using bond pricing models and by shifting each market rate assumption in the model and aggregating their portfolio's sensitivity to these market rates. Often, these factors are combined into scenarios representing expected central bank policies, inflation expectations, and/or anticipated fiscal policy changes. When portfolios are dominated by equities, risk managers typically categorize the equities by broad country markets, industries, and market capitalization levels. Also, they may additionally regress the returns of their portfolios against fundamental factor histories (such as those for growth, value, momentum, and capitalization size) to understand their exposure to such factors.

Portfolios with full fair value accounting (also called mark-to-market accounting), such as US mutual funds, European UCITS funds, and the held-for-sale portfolios of banks, are very well suited to such risk measures as VaR, economic capital (the amount of capital a firm needs to hold if it is to survive severe losses from the risks in its businesses), duration, and beta—all of which rely on measuring the changes in the fair values of assets. Asset/liability gap models are more meaningful when portfolios are subject to book value accounting in whole or in part.

Banks

Banks need to balance a number of sometimes competing aspects of risk to manage their business and meet the expectations of equity investors/equity analysts, bond investors, credit rating agencies, depositors, and regulatory entities. Some banks apply risk measures differently depending on whether the portfolio being assessed is designated as a "held-to-maturity" portfolio, which requires book value accounting, or a "held-for-sale" or "trading book" portfolio, which requires fair value accounting. Other banks will use fair value measures for all risk assessments regardless of the designation used for accounting purposes. In the following list are some of the factors that banks seek to address through their use of risk tools. In compiling this list, we have assumed that banks may treat measures differently depending on accounting treatment.

- *Liquidity gap:* The extent of any liquidity and asset/liability mismatch. The ability to raise sufficient cash for foreseeable payment needs; a view of the liquidity of assets, as well as the expected repayment date of debt.
- *VaR:* The value at risk for the held-for-sale or trading (fair value) portion of the balance sheet.
- *Leverage:* A leverage ratio is typically computed, sometimes according to a regulatory requirement or to an internally determined measure. Leverage ratios will weight risk assets using a variety of methods and rules and divide this weighted asset figure by equity. The result is that riskier assets will be assigned a greater weighting and less risky assets a lower weighting so that more equity is required to support riskier assets.

- *Sensitivities:* For the held-for-sale portion of their balance sheet, banks measure duration, key rate duration or partial duration, and credit spread duration for interest rate risk positions. Banks will also measure foreign exchange exposure and any equity or commodity exposures. All these exposure measures will include the delta sensitivities of options with any other exposures to the same underlying asset and will also monitor gamma and vega exposures of options. Gamma and vega exposures can be broken out by term to identify how much of these risks come from long-dated versus short-dated options.
- *Economic capital:* This is measured by blending the company's market, credit, and operational risk measures to estimate the total loss the company could suffer at a very high level of confidence (e.g., 99% to 99.99%), usually in one year's time. Economic capital measures are applied to the full balance sheet, including both the held-for-sale and held-for-investment portfolios, and include market, credit, and operational risk capital.
- *Scenario analysis:* Stress tests are applied to the full balance sheet and augment economic capital and liquidity; they are used to identify whether capital is sufficient for targeted, strong negative shocks. Outside of stress testing, significant scenario analysis takes places. Scenario analysis is used to examine how the full balance sheet might be affected by different interest rate, inflation, and credit environments, such as unemployment levels for credit card lenders, home price appreciation/depreciation for mortgage lenders, and business cycle stresses for corporate lenders.

It is common for banks to compute risk measures in distinct business units and geographies and then aggregate these measures to the parent company entity.

Asset Managers

Asset managers are not typically regulated with regard to sufficient capital or liquidity; they are more commonly regulated for fair treatment of investors—that disclosures are full and accurate, that marketing is not misleading, that one client is not favored over the other. In some jurisdictions, certain market risk measures may be used to define risk limits for different fund types.

In asset management portfolios, risk management efforts are focused primarily on volatility, probability of loss, or probability of underperforming a benchmark rather than insolvency. A diversified, unleveraged, long-only fund is unlikely to see asset values decline below zero in the absence of a wholesale withdrawal of assets by the firm's clients. Although service costs and other items make insolvency a technical possibility, in practice, insolvency is a much higher threat for leveraged portfolios. Although derivatives use by asset managers can create effective leverage, these positions are often balanced by an amount of cash in the portfolio equal to the notional exposure created by the derivatives mitigating, if not fully eliminating, the impact of leverage.

Asset managers typically measure and view each portfolio separately with respect to its own constraints and limits. However, there are a few exceptions:

- Long-only asset managers: If the adviser has invested its own capital in any of the funds that it manages, these investments may need to be aggregated for the firm to assess its risk exposures across portfolios.
- Hedge funds: A hedge fund manager needs to aggregate the adviser's side-by-side investment in the various funds it advises.
- Funds of funds: Risk measures for these portfolios typically aggregate the risks of the underlying hedge funds to the master fund level.

An asset manager may choose to aggregate exposures across all funds and strategies to determine if there are unusual concentrations in individual securities or counterparties that would make management actions across all portfolios difficult to carry out (e.g., a single portfolio's holdings in a given security may not pose a liquidity risk, but if the firm were to aggregate all of its holdings in that security, it may find that the portfolio fails to meet the desired liquidity target).

It is important when observing risk measures for asset managers to determine whether the measures represent the backward-looking variability of realized returns in the portfolio as it was then constituted or use the current portfolio and measure its potential loss. Backward-looking returns-based measures (typically including standard deviation, *ex post* tracking error, Sharpe ratio, information ratio, and historical beta) have the value of showing the fund's behavior over time and help assess the skill of the manager. Only an analysis of the current holdings, however, will reveal current risk exposures. Measures that use current holdings typically include VaR, *ex ante* tracking error, duration and forward-looking beta, stress tests, and scenario analyses. All risk and performance measures can be conducted on past portfolio holdings or current portfolio holdings; it is important for the user of any measure to determine which ingredients (which set of portfolio holdings, and for market history, what length and smoothing techniques) have been used in order to use it correctly. Assessing the trends in risk exposures, including whether risk has recently risen or if other important changes have taken place in the strategy, can be accomplished by tracking the risk measures through time.

Traditional Asset Managers

Asset managers that use little leverage typically find relative risk measures most meaningful and actionable. The decision to invest in a given asset class is normally the client's, not the adviser's. The adviser seeks to outperform the benchmark representative of the asset class. Exceptions include absolute return funds and asset allocation strategies, but even these can be measured relative to a benchmark. For absolute return strategies, the benchmark is typically cash or a cash-like hurdle rate. When cash is the benchmark, VaR and *ex ante* tracking error will be effectively the same if measured using the same holding period and confidence interval. (Cash has no volatility, so adding a cash benchmark into a relative VaR calculation does not affect the calculation because its zero volatility cancels out its impact; thus, the resulting calculation is the same as the VaR of the portfolio.) Asset allocation funds can use an asset allocation index as the benchmark for a relative risk measure, or they can use a custom combination of market benchmarks.

Although banks, insurers, and other market participants favor measuring VaR in currency terms relevant for the institution (e.g., dollars for a US-based insurer, yen for a Japanese bank) and measure duration and similar statistics as the value change for a 1 bp interest rate change, long-only asset managers generally prefer to express VaR in percentage terms and will divide VaR and duration by the net assets of the portfolio being analyzed. (Note that using returns as the fundamental source of data removes the last step in calculating VaR: multiplying by the size of the portfolio.)

A typical sample of risk measures used by asset managers includes the following:

- *Position limits:* Asset managers use position limits as the most frequent form of risk control for the portfolios they manage, particularly in fund offering documents that need to be understandable to a broad range of investors. Position limits include restrictions on country, currency, sector, and asset class. They may measure them in absolute terms or relative to a benchmark, and they are almost always expressed as a percentage of the portfolio's value.

- *Sensitivities:* Asset managers use the full range of sensitivity measures, including option-adjusted duration, key rate duration, and credit spread duration, and they will typically include the delta exposure of options in these measures. Measures can be expressed in absolute terms as well as relative to a benchmark.
- *Beta sensitivity:* Beta is frequently used for equity-only accounts.
- *Liquidity:* Asset managers often look at the liquidity characteristics of the assets in their portfolios. For equity portfolios, it is common to measure what percentage of daily average trading volume the portfolio holds of each equity security and how many days it would take to liquidate a security if the manager did not want it to be too large a portion of trading volume to avoid taking a price concession.
- *Scenario analysis:* Long-only asset managers typically use stress tests or scenario analyses to verify that the risks in the portfolio are as they have been disclosed to investors and to identify any unusual behavior that could arise in stressed markets.
- *Redemption risk:* Open-end fund managers often assess what percentage of the portfolio could be redeemed at peak times and track this behavior across the funds and asset classes they manage.
- *Ex post versus ex ante tracking error:* Limits on *ex ante* tracking error are often used by traditional asset managers as a key risk metric for the portfolios they manage. It provides an estimate of the degree to which the current portfolio could underperform its benchmark. It is worth noting the distinction between *ex post* tracking error and *ex ante* tracking error: Asset managers use *ex post* tracking error to identify sources of performance and manager skill and *ex ante* tracking error to identify whether today's positions could give rise to unexpected potential performance. *Ex post* tracking error measures the historical deviation between portfolio returns and benchmark returns, and thus both the portfolio holdings and market returns are historical in this measure. *Ex ante* tracking error takes today's benchmark-relative position and exposes it to the variability of past markets to estimate what kind of benchmark-relative performance could arise from the current portfolio. *Ex post* tracking error is a useful tool for assessing manager skill and behavior. The day after a large change in portfolio strategy, *ex ante* tracking will immediately reflect the portfolio's new return profile, whereas *ex post* tracking error will not do so until the new strategy has been in place long enough to dominate the data history. (If *ex post* tracking error is computed using 200 days of history, the day after a large strategy change, only 1 of the 200 data points will reflect the current risk positioning.) Some asset managers focus on maintaining *ex ante* tracking error boundaries for the portfolios they manage to monitor and balance the potential performance impact of the active risks they are taking. **Active share** is a measure of that percentage of the portfolio that differs from the benchmark (i.e., a deviation from the benchmark). It is often monitored to help limit tracking error of the portfolio.
- *VaR:* VaR is less commonly used as a risk measure than *ex ante* tracking error by traditional asset managers, but it is used by some—particularly for portfolios that are characterized as "absolute return" strategies for which a given market benchmark may not serve as the portfolio objective.

Hedge Funds

Similar to banks, hedge funds that use leverage need to observe sources and uses of cash through time, including when credit lines could be withdrawn, and need to simulate the interplay between market movements, margin calls, and the redemption rights of investors in order to understand worst-case needs for cash. A sample of the typical range of hedge fund market risk measures includes the following:

- *Sensitivities:* All hedge fund strategies will display some form of sensitivity or exposure, so the full range of sensitivity measures are useful for hedge fund risk management.
- *Gross exposure:* Long–short, market neutral, and arbitrage strategies will typically measure long exposure, short exposure, and gross exposure (the sum of the absolute value of long plus short positions) separately. Gross position risk is an important guide to the importance of correlation risk for the portfolio.
- *Leverage:* Leverage measures are common for hedge funds. It is important to understand how the measure is treating derivatives and what elements appear in the numerator versus the denominator because there are many different ways to execute the measure.
- *VaR:* Hedge funds that use VaR measures tend to focus on high confidence intervals (more than 90%) and short holding periods, and they rarely use a benchmark-relative measure.
- *Scenarios:* Hedge funds commonly use scenario/stress tests that are well tuned to the specific risks of their strategy—in merger arbitrage strategies, for example, the chance that the merger will not take place.
- *Drawdown:* In the case of the following types of hedge fund strategies, standard deviation and historical beta measures can be particularly misleading when seeking to understand what the more extreme risks can be. This is because the strategies listed frequently display decidedly non-normal return distributions, and when this is true, standard deviation is not a good guide to worst-case outcomes. For the following strategies, any historical standard deviation or historical beta measures should be supplemented by a measure of what has been the **maximum drawdown**, often defined as the worst-returning month or quarter for the portfolio or the worst peak-to-trough decline in a portfolio's returns:
    - Strategies that focus on credit risk taking, such as long–short credit, credit arbitrage, or bankruptcy investing
    - Strategies that focus on events, such as merger arbitrage
    - Strategies that make meaningful investments in non-publicly issued assets or other assets that do not reliably have a daily, independent fair value determination
    - Strategies that invest in illiquid asset classes or take large positions relative to market size in any asset class
    - Strategies that sell options or purchase bonds with embedded options
    - Strategies that are highly reliant on correlation relationships, such as equity market neutral

In addition, it is not uncommon for those investing in hedge funds to look at the returns of the hedge fund during a relevant historical period, such as the 2008 financial crisis.

PENSION FUNDS AND INSURERS

12

☐ describe risk measures used by banks, asset managers, pension funds, and insurers

A defined benefit pension plan is required to make payments to its pensioners in the future that are typically determined as a function of a retiree's final salary. This differs from a defined contribution plan, in which the plan's sponsor may be required to make contributions currently but is not responsible to ensure that they grow to a particular future amount. To meet the required payouts, defined benefit plans have significant market risk management responsibilities. This section describes the practices of defined benefit pension plans only; all mentions in this section of "pension funds" or "pension plans" refer to defined benefit pensions.

The risk management goal for pension funds is to be sufficiently funded to make future payments to pensioners. The requirements for sufficient funding vary from country to country. Different jurisdictions will have regulations concerning such items as how to compute the present value of pension liabilities (including which interest rates are permitted to be used as a discount rate) and what the sponsor of the pension plan is required to contribute when the assets in the pension fund are lower than the present value of the liabilities. In addition, some jurisdictions impose taxes when surplus—the value of the assets less the value of the liabilities—is withdrawn for other use by the plan sponsor. Although these regional differences will shape the practice of pension plan risk management in different countries, it is typically an exercise in ensuring that the plan is not likely to become significantly under- or overfunded. Overfunding occurs when the funding ratio (the assets divided by the present value of the liabilities) is greater than 100%; underfunding occurs when the funding ratio is under 100%. Overfunding may be cured over time by the plan sponsor not needing to make regular contributions to the plan because the number of employees and their salary levels, which drive the pension benefit, are growing. Underfunding, if not cured by growth in the assets in the fund over a suitable time horizon as permitted by regulation, is cured by the plan sponsor contributing to the fund. The pension plan's actions will also vary depending on its age (whether it is a new or established plan) and whether it is currently meaningfully under- or overfunded. Important market risk measures or methods for pension funds often include the following:

- *Interest rate and curve risk:* The first step of risk measurement for pension funds is the analysis of expected payments to pensioners in the future. The expected future cash flows are grouped by maturity. In the case of an international pension fund that must make future payouts in multiple currencies, they may also be grouped by currency. In cases in which the jurisdiction requires a particular fixed-income instrument or curve be used to provide the discount rate for arriving at the present value of the pension liability (such as corporate bonds in the United States, inflation-linked gilts in the United Kingdom, or government bonds in the Netherlands), the liability cash flows will be expressed as a short position at the relevant points on the curve.
- *Surplus at risk:* This measure is an application of VaR. It is computed by entering the assets in the portfolio into a VaR model as long positions and the pension liabilities as short fixed-income positions. It estimates how much the assets might underperform the liabilities, usually over one year, and pension plan sponsors may vary with respect to how high a level of confidence they choose to use (e.g., 84%, 95%, 99%). If the assets in the portfolio

were invested precisely in the same fixed-income instruments to which the liabilities have been apportioned and in the same amounts, it would result in zero surplus at risk. In practice, however, it may be impossible to invest in the sizes required in the particular fixed-income instruments specified in the liability analysis, so the pension will invest in other, non-fixed-income investments, such as equities or real assets. The more volatile the investments in the pension fund and the less well correlated these assets are with the liabilities, the higher the surplus at risk. The pension fund may set a threshold level or limit on surplus at risk; when the pension fund's surplus at risk exceeds this limit, pension staff will change the fund's asset allocation to make the assets in the fund better match the liabilities. This liability-focused form of pension investing is commonly referred to as "liability driven investing."

- *Liability hedging exposures versus return generating exposures:* Although matching liabilities is an important goal of pension fund management, it is not the only goal. Pension staff may separate their investment portfolio into investments designed to match the pension liability versus those meant to generate excess returns. The precise instruments linked to the liability cannot always be directly invested in, so a separate portion of the portfolio may be necessary and should perform the function of earning returns that can minimize the chance of having an over- or underfunded status greater than the pension fund's risk tolerance. The return-generating portion of the portfolio also helps to hedge the potential for future changes in the size of the liability that could be caused by longevity risk or by wage growth that exceeds the forecasts currently used to compute the liability.

Insurers

Insurers in the largest global economies are subject to significant regulation and accounting oversight regarding how they must retain reserves and reflect their liabilities. Regulation may also affect the pricing permitted by product line. It is common for insurers to aggregate risk from underlying business units to arrive at a firm-wide view of risk.

Insurance liabilities vary in their correlation with financial markets. The risk metrics of property and casualty insurance differ significantly from those used for life insurance and annuity products. Property and casualty insurance, including home, auto, corporate liability insurance, and health insurance, are typically not highly correlated with financial asset markets.

Insurers focus on managing a number of forms of insurance risk, for which they may use such tools as reinsurance and geographic dispersion. The market risk management measures in the property and casualty lines of business include the following:

- *Sensitivities and exposures:* Insurers often design an asset allocation for these portfolios and monitor current exposures to remain within the target ranges set forth in the target asset allocation.
- *Economic capital and VaR:* The risk measurement focus for these lines of business is capital at risk and VaR. The premiums earned in these areas are typically set to compensate for the expected payouts (usually defined as a range of possible payouts), so it is only in cases of greater-than-expected payouts that capital is tapped. The risk modeling effort is to estimate what that catastrophic loss amount could be at a given level of probability.

Assessment of the risk to economic capital will include the market risks in the portfolio as well as characteristics of the insurance exposures and reinsurance coverage.

- *Scenario analysis:* Insurers use scenario analysis like other market participants that have capital at risk, such as banks and hedge funds. For the property and casualty lines, these scenarios may stress the market risks and the insurance risks in the same scenario.

Insurers do not focus on matching assets with liabilities in their property and casualty lines of business. Investment portfolios are not designed to pay out insurance claims in property and casualty insurance businesses; the premium income is primarily used for that purpose. These investments are designed to achieve a good absolute return within the constraints imposed under regulatory reserve requirements. Riskier assets are discounted relative to safer, fixed-income assets in measuring required reserves.

Life insurance and annuities have stronger ties to the financial markets, even while retaining distinct mortality-based risk profiles. Life liabilities are very long, and the reserves that insurers are required to maintain by insurance regulators are highly dependent on discount rate assumptions. Non-financial inputs include assumptions about mortality and which policyholders will either tap into options in their policy to add coverage at a given level or cancel their policy. Annuities produce returns based on financial assets, with some extra optionality driven by any life insurance elements embedded in the policy. These activities are paired with long-term investment portfolios in a variety of assets that are designed to help the insurer meet future claims.

For life portfolios, market risk measures include the following:

- *Sensitivities:* The exposures of the investment portfolio and the annuity liability are measured and monitored.
- *Asset and liability matching:* The investment portfolio is not designed to be a perfect match to the liabilities, but it is more closely matched to liabilities than is the case in property and casualty insurance.
- *Scenario analysis:* The main focus of risk measurement for the life lines of insurance are measures of potential stress losses based on the differences between the assets in which the insurance company has invested and the liabilities driven by the insurance contracts it has written to its customers. Scenario analyses need to stress both market and non-market sources of cash flow change (in which non-market changes can include changes in longevity).

EXAMPLE 11

Uses of Risk Measures by Market Participants

1. Which type of market participant is *most likely* to consistently express risk measures as a percentage of assets and relative to a benchmark?

   A. Banks

   B. Corporations

   C. Long-only asset managers

Solution:

C is correct. Long-only asset managers most commonly express risk measures in percentage terms and relative to a benchmark, whereas the entities in answers A and B measure risk more commonly in currency units and in absolute terms (not relative to a benchmark). Banks occasionally express risk measures, such as economic capital, as a percentage of assets or other balance sheet measures, but bank risk measures are typically expressed in currency units.

2. How does *ex ante* tracking error differ from *ex post* tracking error?

   A. *Ex ante* tracking error takes into account the behavior of options, whereas *ex post* tracking error does not.

   B. *Ex post* tracking error uses a more accurate forecast of future markets than the forecast used for *ex ante* tracking error.

   C. *Ex ante* tracking error uses *current* portfolio holdings exposed to the variability of historical markets, whereas *ex post* tracking error measures the variability of *historical* portfolio holdings in historical markets.

Solution:

C is correct. A is incorrect because although *ex post* tracking error accounts for the options that were in the portfolio in the past, *ex ante* tracking error might actually misstate the risk of options if it is computed using the parametric method. B is incorrect because *ex post* tracking error is not aiming to forecast the future; it is only measuring the variability of past results.

SUMMARY

This reading on market risk management models covers various techniques used to manage the risk arising from market fluctuations in prices and rates. The key points are summarized as follows:

- Value at risk (VaR) is the minimum loss in either currency units or as a percentage of portfolio value that would be expected to be incurred a certain percentage of the time over a certain period of time given assumed market conditions.
- VaR requires the decomposition of portfolio performance into risk factors.
- The three methods of estimating VaR are the parametric method, the historical simulation method, and the Monte Carlo simulation method.
- The parametric method of VaR estimation typically provides a VaR estimate from the left tail of a normal distribution, incorporating the expected returns, variances, and covariances of the components of the portfolio.
- The parametric method exploits the simplicity of the normal distribution but provides a poor estimate of VaR when returns are not normally distributed, as might occur when a portfolio contains options.
- The historical simulation method of VaR estimation uses historical return data on the portfolio's current holdings and allocation.
- The historical simulation method has the advantage of incorporating events that actually occurred and does not require the specification of a distribution or the estimation of parameters, but it is only useful to the extent that the future resembles the past.
- The Monte Carlo simulation method of VaR estimation requires the specification of a statistical distribution of returns and the generation of random outcomes from that distribution.

- The Monte Carlo simulation method is extremely flexible but can be complex and time consuming to use.
- There is no single right way to estimate VaR.
- The advantages of VaR include the following: It is a simple concept; it is relatively easy to understand and easily communicated, capturing much information in a single number. It can be useful in comparing risks across asset classes, portfolios, and trading units and, as such, facilitates capital allocation decisions. It can be used for performance evaluation and can be verified by using backtesting. It is widely accepted by regulators.
- The primary limitations of VaR are that it is a subjective measure and highly sensitive to numerous discretionary choices made in the course of computation. It can underestimate the frequency of extreme events. It fails to account for the lack of liquidity and is sensitive to correlation risk. It is vulnerable to trending or volatility regimes and is often misunderstood as a worst-case scenario. It can oversimplify the picture of risk and focuses heavily on the left tail.
- There are numerous variations and extensions of VaR, including conditional VaR (CVaR), incremental VaR (IVaR), and marginal VaR (MVaR), that can provide additional useful information.
- Conditional VaR is the average loss conditional on exceeding the VaR cutoff.
- Incremental VaR measures the change in portfolio VaR as a result of adding or deleting a position from the portfolio or if a position size is changed relative to the remaining positions.
- MVaR measures the change in portfolio VaR given a small change in the portfolio position. In a diversified portfolio, MVaRs can be summed to determine the contribution of each asset to the overall VaR.
- *Ex ante* tracking error measures the degree to which the performance of a given investment portfolio might deviate from its benchmark.
- Sensitivity measures quantify how a security or portfolio will react if a single risk factor changes. Common sensitivity measures are beta for equities; duration and convexity for bonds; and delta, gamma, and vega for options. Sensitivity measures do not indicate which portfolio has greater loss potential.
- Risk managers can use deltas, gammas, vegas, durations, convexities, and betas to get a comprehensive picture of the sensitivity of the entire portfolio.
- Stress tests apply extreme negative stress to a particular portfolio exposure.
- Scenario measures, including stress tests, are risk models that evaluate how a portfolio will perform under certain high-stress market conditions.
- Scenario measures can be based on actual historical scenarios or on hypothetical scenarios.
- Historical scenarios are scenarios that measure the portfolio return that would result from a repeat of a particular period of financial market history.
- Hypothetical scenarios model the impact of extreme movements and co-movements in different markets that have not previously occurred.
- Reverse stress testing is the process of stressing the portfolio's most significant exposures.
- Sensitivity and scenario risk measures can complement VaR. They do not need to rely on history, and scenarios can be designed to overcome an assumption of normal distributions.

- Limitations of scenario measures include the following: Historical scenarios are unlikely to re-occur in exactly the same way. Hypothetical scenarios may incorrectly specify how assets will co-move and thus may get the magnitude of movements wrong. And, it is difficult to establish appropriate limits on a scenario analysis or stress test.
- Constraints are widely used in risk management in the form of risk budgets, position limits, scenario limits, stop-loss limits, and capital allocation.
- Risk budgeting is the allocation of the total risk appetite across sub-portfolios.
- A scenario limit is a limit on the estimated loss for a given scenario, which, if exceeded, would require corrective action in the portfolio.
- A stop-loss limit either requires a reduction in the size of a portfolio or its complete liquidation (when a loss of a particular size occurs in a specified period).
- Position limits are limits on the market value of any given investment.
- Risk measurements and constraints in and of themselves are not restrictive or unrestrictive; it is the limits placed on the measures that drive action.
- The degree of leverage, the mix of risk factors to which the business is exposed, and accounting or regulatory requirements influence the types of risk measures used by different market participants.
- Banks use risk tools to assess the extent of any liquidity and asset/liability mismatch, the probability of losses in their investment portfolios, their overall leverage ratio, interest rate sensitivities, and the risk to economic capital.
- Asset managers' use of risk tools focuses primarily on volatility, probability of loss, or the probability of underperforming a benchmark.
- Pension funds use risk measures to evaluate asset/liability mismatch and surplus at risk.
- Property and casualty insurers use sensitivity and exposure measures to ensure exposures remain within defined asset allocation ranges. They use economic capital and VaR measures to estimate the impairment in the event of a catastrophic loss. They use scenario analysis to stress the market risks and insurance risks simultaneously.
- Life insurers use risk measures to assess the exposures of the investment portfolio and the annuity liability, the extent of any asset/liability mismatch, and the potential stress losses based on the differences between the assets in which they have invested and the liabilities resulting from the insurance contracts they have written.

REFERENCES

Malkiel, Burton. 2007. *A Random Walk Down Wall Street*. New York: W.W. Norton.

PRACTICE PROBLEMS

The following information relates to questions 1-5

Randy Gorver, chief risk officer at Eastern Regional Bank, and John Abell, assistant risk officer, are currently conducting a risk assessment of several of the bank's independent investment functions. These reviews include the bank's fixed-income investment portfolio and an equity fund managed by the bank's trust department. Gorver and Abell are also assessing Eastern Regional's overall risk exposure.

Eastern Regional Bank Fixed-Income Investment Portfolio

The bank's proprietary fixed-income portfolio is structured as a barbell portfolio: About half of the portfolio is invested in zero-coupon Treasuries with maturities in the 3- to 5-year range (Portfolio P_1), and the remainder is invested in zero-coupon Treasuries with maturities in the 10- to 15-year range (Portfolio P_2). Georges Montes, the portfolio manager, has discretion to allocate between 40% and 60% of the assets to each maturity "bucket." He must remain fully invested at all times. Exhibit 1 shows details of this portfolio.

Exhibit 1: US Treasury Barbell Portfolio

	Maturity	
	P_1	P_2
	3–5 Years	10–15 Years
Average duration	3.30	11.07
Average yield to maturity	1.45%	2.23%
Market value	$50.3 million	$58.7 million

Trust Department's Equity Fund

a. **Use of Options:** The trust department of Eastern Regional Bank manages an equity fund called the Index Plus Fund, with $325 million in assets. This fund's objective is to track the S&P 500 Index price return while producing an income return 1.5 times that of the S&P 500. The bank's chief investment officer (CIO) uses put and call options on S&P 500 stock index futures to adjust the risk exposure of certain client accounts that have an investment in this fund. The portfolio of a 60-year-old widow with a below-average risk tolerance has an investment in this fund, and the CIO has asked his assistant, Janet Ferrell, to propose an options strategy to bring the portfolio's delta to 0.90.

b. **Value at Risk**: The Index Plus Fund has a value at risk (VaR) of $6.5 million at 5% for one day. Gorver asks Abell to write a brief summary of the portfolio VaR for the report he is preparing on the fund's risk position.

Combined Bank Risk Exposures

The bank has adopted a new risk policy, which requires forward-looking risk assessments in addition to the measures that look at historical risk characteristics. Management has also become very focused on tail risk since the subprime crisis and is evaluating the bank's capital allocation to certain higher-risk lines of business. Gorver must determine what additional risk metrics to include in his risk reporting to address the new policy. He asks Abell to draft a section of the risk report that will address the risk measures' adequacy for capital allocation decisions.

1. If Montes is expecting a 50 bp increase in yields at all points along the yield curve, which of the following trades is he *most likely* to execute to minimize his risk?
   A. Sell $35 million of P_2 and reinvest the proceeds in three-year bonds
   B. Sell $15 million of P_2 and reinvest the proceeds in three-year bonds
   C. Reduce the duration of P_2 to 10 years and reduce the duration of P_1 to 3 years

2. Which of the following options strategies is Ferrell *most likely* to recommend for the client's portfolio?
   A. Long calls
   B. Short calls
   C. Short puts

3. Which of the following statements regarding the VaR of the Index Plus Fund is correct?
   A. The expected maximum loss for the portfolio is $6.5 million.
   B. Five percent of the time, the portfolio can be expected to experience a loss of at least $6.5 million.
   C. Ninety-five percent of the time, the portfolio can be expected to experience a one-day loss of no more than $6.5 million.

4. To comply with the new bank policy on risk assessment, which of the following is the *best* set of risk measures to add to the chief risk officer's risk reporting?
   A. Conditional VaR, stress test, and scenario analysis
   B. Monte Carlo VaR, incremental VaR, and stress test
   C. Parametric VaR, marginal VaR, and scenario analysis

5. Which of the following statements should *not* be included in Abell's report to management regarding the use of risk measures in capital allocation decisions?
   A. VaR measures capture the increased liquidity risk during stress periods.

B. Stress tests and scenario analysis can be used to evaluate the effect of outlier events on each line of business.

C. VaR approaches that can accommodate a non-normal distribution are critical to understand relative risk across lines of business.

The following information relates to questions 6-13

Tina Ming is a senior portfolio manager at Flusk Pension Fund (Flusk). Flusk's portfolio is composed of fixed-income instruments structured to match Flusk's liabilities. Ming works with Shrikant McKee, Flusk's risk analyst.

Ming and McKee discuss the latest risk report. McKee calculated value at risk (VaR) for the entire portfolio using the historical method and assuming a look-back period of five years and 250 trading days per year. McKee presents VaR measures in Exhibit 1.

Exhibit 1: Flusk Portfolio VaR (in $ millions)

Confidence Interval	Daily VaR	Monthly VaR
95%	1.10	5.37

After reading McKee's report, Ming asks why the number of daily VaR breaches over the last year is zero even though the portfolio has accumulated a substantial loss.

Next, Ming requests that McKee perform the following two risk analyses on Flusk's portfolio:

Analysis 1	Use scenario analysis to evaluate the impact on risk and return of a repeat of the last financial crisis.
Analysis 2	Estimate over one year, with a 95% level of confidence, how much Flusk's assets could underperform its liabilities.

Ming recommends purchasing newly issued emerging market corporate bonds that have embedded options. Prior to buying the bonds, Ming wants McKee to estimate the effect of the purchase on Flusk's VaR. McKee suggests running a stress test using a historical period specific to emerging markets that encompassed an extreme change in credit spreads.

At the conclusion of their conversation, Ming asks the following question about risk management tools: "What are the advantages of VaR compared with other risk measures?"

6. Based on Exhibit 1, Flusk's portfolio is expected to experience:

A. a minimum daily loss of $1.10 million over the next year.

B. a loss over one month equal to or exceeding $5.37 million 5% of the time.

C. an average daily loss of $1.10 million 5% of the time during the next 250 trading days.

7. The number of Flusk's VaR breaches *most likely* resulted from:
   A. using a standard normal distribution in the VaR model.
   B. using a 95% confidence interval instead of a 99% confidence interval.
   C. lower market volatility during the last year compared with the lookback period.

8. To perform Analysis 1, McKee should use historical bond:
   A. prices.
   B. yields.
   C. durations.

9. The limitation of the approach requested for Analysis 1 is that it:
   A. omits asset correlations.
   B. precludes incorporating portfolio manager actions.
   C. assumes no deviation from historical market events.

10. The estimate requested in Analysis 2 is *best* described as:
    A. liquidity gap.
    B. surplus at risk.
    C. maximum drawdown.

11. Which measure should McKee use to estimate the effect on Flusk's VaR from Ming's portfolio recommendation?
    A. Relative VaR
    B. Incremental VaR
    C. Conditional VaR

12. When measuring the portfolio impact of the stress test suggested by McKee, which of the following is *most likely* to produce an accurate result?
    A. Marginal VaR
    B. Full revaluation of securities
    C. The use of sensitivity risk measures

13. The risk management tool referenced in Ming's question:
    A. is widely accepted by regulators.
    B. takes into account asset liquidity.
    C. usually incorporates right-tail events.

The following information relates to questions 14-20

Carol Kynnersley is the chief risk officer at Investment Management Advisers (IMA). Kynnersley meets with IMA's portfolio management team and investment advisers to discuss the methods used to measure and manage market risk and how risk metrics are presented in client reports.

The three most popular investment funds offered by IMA are the Equity Opportunities, the Diversified Fixed Income, and the Alpha Core Equity. The Equity Opportunities Fund is composed of two exchange-traded funds: a broadly diversified large-cap equity product and one devoted to energy stocks. Kynnersley makes the following statements regarding the risk management policies established for the Equity Opportunities portfolio:

Statement 1 IMA's preferred approach to model value at risk (VaR) is to estimate expected returns, volatilities, and correlations under the assumption of a normal distribution.

Statement 2 In last year's annual client performance report, IMA stated that a hypothetical $6 million Equity Opportunities Fund account had a daily 5% VaR of approximately 1.5% of portfolio value.

Kynnersley informs the investment advisers that the risk management department recently updated the model for estimating the Equity Opportunities Fund VaR based on the information presented in Exhibit 1.

Exhibit 1: Equity Opportunities Fund—VaR Model Input Assumptions

	Large-Cap ETF	Energy ETF	Total Portfolio
Portfolio weight	65.0%	35.0%	100.0%
Expected annual return	12.0%	18.0%	14.1%
Standard deviation	20.0%	40.0%	26.3%
Correlation between ETFs: 0.90			
Number of trading days/year: 250			

For clients interested in fixed-income products, IMA offers the Diversified Fixed-Income Fund. Kynnersley explains that the portfolio's bonds are all subject to interest rate risk. To demonstrate how fixed-income exposure measures can be used to identify and manage interest rate risk, Kynnersley distributes two exhibits featuring three hypothetical Treasury coupon bonds (Exhibit 2) under three interest rate scenarios (Exhibit 3).

Exhibit 2: Fixed-Income Risk Measure

Hypothetical Bond	Duration
Bond 1	1.3
Bond 2	3.7
Bond 3	10.2

Exhibit 3: Interest Rate Scenarios

Scenario	Interest Rate Environment
Scenario 1	Rates increase 25 bps
Scenario 2	Rates increase 10 bps
Scenario 3	Rates decrease 20 bps

One of the investment advisers comments that a client recently asked about the performance of the Diversified Fixed-Income Fund relative to its benchmark, a broad fixed-income index. Kynnersley informs the adviser as follows:

Statement 3 The Diversified Fixed-Income Fund manager monitors the historical deviation between portfolio returns and benchmark returns. The fund prospectus stipulates a target deviation from the benchmark of no more than 5 bps.

Kynnersley concludes the meeting by reviewing the constraints IMA imposes on securities included in the Alpha Core Equity Fund. The compliance department conducts daily oversight using numerous risk screens and, when indicated, notifies portfolio managers to make adjustments. Kynnersley makes the following statement:

Statement 4 It is important that all clients investing in the fund be made aware of IMA's compliance measures. The Alpha Core Equity Fund restricts the exposure of individual securities to 1.75% of the total portfolio.

14. Based on Statement 1, IMA's VaR estimation approach is *best* described as the:

A. parametric method.

B. historical simulation method.

C. Monte Carlo simulation method.

15. In Statement 2, Kynnersley implies that the portfolio:

A. is at risk of losing $4,500 each trading day.

B. value is expected to decline by $90,000 or more once in 20 trading days.

C. has a 5% chance of falling in value by a maximum of $90,000 on a single trading day.

16. Based *only* on Statement 2, the risk measurement approach:

A. ignores right-tail events in the return distribution.

B. is similar to the Sharpe ratio because it is backward looking.

C. provides a relatively accurate risk estimate in both trending and volatile regimes.

17. Based on Exhibit 1, the daily 5% VaR estimate is *closest* to:

A. 1.61%.

B. 2.42%.

C. 2.69%.

18. Based *only* on Exhibits 2 and 3, it is *most likely* that under:

A. Scenario 1, Bond 2 outperforms Bond 1.

B. Scenario 2, Bond 1 underperforms Bond 3.

C. Scenario 3, Bond 3 is the best performing security.

19. The risk measure referred to in Statement 3 is:

A. active share.

B. beta sensitivity

C. *ex post* tracking error.

20. In Statement 4, Kynnersley describes a constraint associated with a:

A. risk budget.

B. position limit.

C. stop-loss limit.

The following information relates to questions 21-26

Hiram Life (Hiram), a large multinational insurer located in Canada, has received permission to increase its ownership in an India-based life insurance company, LICIA, from 26% to 49%. Before completing this transaction, Hiram wants to complete a risk assessment of LICIA's investment portfolio. Judith Hamilton, Hiram's chief financial officer, has been asked to brief the management committee on investment risk in its India-based insurance operations.

LICIA's portfolio, which has a market value of CAD260 million, is currently structured as shown in Exhibit 1. Despite its more than 1,000 individual holdings, the portfolio is invested predominantly in India. The Indian government bond market is highly liquid, but the country's mortgage and infrastructure loan markets, as well as the corporate bond market, are relatively illiquid. Individual mortgage and corporate bond positions are large relative to the normal trading volumes in these securities. Given the elevated current and fiscal account deficits, Indian investments are also subject to above-average economic risk.

Hamilton begins with a summary of the India-based portfolio. Exhibit 1 presents the current portfolio composition and the risk and return assumptions used to estimate value at risk (VaR).

Exhibit 1: Selected Assumptions for LICIA's Investment Portfolio

	Allocation	Average Daily Return	Daily Standard Deviation
India government securities	50%	0.015%	0.206%
India mortgage/infrastructure loans	25%	0.045%	0.710%
India corporate bonds	15%	0.025%	0.324%
India equity	10%	0.035%	0.996%

Infrastructure is a rapidly growing asset class with limited return history; the first infrastructure loans were issued just 10 years ago.

Hamilton's report to the management committee must outline her assumptions and provide support for the methods she used in her risk assessment. If needed, she will also make recommendations for rebalancing the portfolio to ensure its risk profile is aligned with that of Hiram.

Hamilton develops the assumptions shown in Exhibit 2, which will be used for estimating the portfolio VaR.

Exhibit 2: VaR Input Assumptions for Proposed CAD260 Million Portfolio

Method	Average Return Assumption	Standard Deviation Assumption
Monte Carlo simulation	0.026%	0.501%
Parametric approach	0.026%	0.501%
Historical simulation	0.023%	0.490%

Hamilton elects to apply a one-day, 5% VaR limit of CAD2 million in her risk assessment of LICIA's portfolio. This limit is consistent with the risk tolerance the committee has specified for the Hiram portfolio.

The markets' volatility during the last 12 months has been significantly higher than the historical norm, with increased frequency of large daily losses, and Hamilton expects the next 12 months to be equally volatile.

She estimates the one-day 5% portfolio VaR for LICIA's portfolio using three different approaches:

Exhibit 3: VaR Results over a One-Day Period for Proposed Portfolio

Method	5% VaR
Monte Carlo simulation	CAD2,095,565
Parametric approach	CAD2,083,610
Historical simulation	CAD1,938,874

The committee is likely to have questions in a number of key areas—the limitations of the VaR report, potential losses in an extreme adverse event, and the reliability of the VaR numbers if the market continues to exhibit higher-than-normal volatility. Hamilton wants to be certain that she has thoroughly evaluated the risks inherent in the LICIA portfolio and compares them with the risks in Hiram's present portfolio.

Hamilton believes the possibility of a ratings downgrade on Indian sovereign debt is high and not yet fully reflected in securities prices. If the rating is lowered, many of the portfolio's holdings will no longer meet Hiram's minimum ratings requirement. A downgrade's effect is unlikely to be limited to the government bond portfolio. All asset classes can be expected to be affected to some degree. Hamilton plans to include a scenario analysis that reflects this possibility to ensure that management has the broadest possible view of the risk exposures in the India portfolio.

21. Given Hamilton's expectations, which of the following models is *most appropriate* to use in estimating portfolio VaR?

    A. Parametric method

    B. Historical simulation method

    C. Monte Carlo simulation method

22. Which risk measure is Hamilton *most likely* to present when addressing the committee's concerns regarding potential losses in extreme stress events?

    A. Relative VaR

    B. Incremental VaR

    C. Conditional VaR

23. The scenario analysis that Hamilton prepares for the committee is *most likely* a:

    A. stress test.

    B. historical scenario.

    C. hypothetical scenario.

24. The scenario analysis that Hamilton prepares for the committee is a valuable tool to supplement VaR *because* it:

    A. incorporates historical data to evaluate the risk in the tail of the VaR distribution.

    B. enables Hamilton to isolate the risk stemming from a single risk factor—the ratings downgrade.

    C. allows the committee to assess the effect of low liquidity in the event of a ratings downgrade.

25. Using the data in Exhibit 2, the portfolio's annual 1% parametric VaR is *closest* to:

    A. CAD17 million.

    B. CAD31 million.

    C. CAD48 million.

26. What additional risk measures would be most appropriate to add to Hamilton's risk assessment?

    A. Delta

B. Duration

C. Tracking error

SOLUTIONS

1. B is correct. Duration is a measure of interest rate risk. To reduce risk in anticipation of an increase in interest rates, Montes would seek to shorten the portfolio's duration. He is limited, however, in the amount he can shift from P_2 to P_1. Selling \$15 million of P_2 reduces that portfolio to the lower end of the permitted 40% to 60% range. By reinvesting the proceeds at the shortest maturities allowed, Montes substantially reduces the portfolio duration.

2. B is correct. An index-tracking portfolio without options has a delta of 1. To achieve a delta of 0.9, the delta of the options position must be negative. Of the three choices, only short calls have a negative delta. Long call options have deltas ranging from 0 to 1. Short calls, therefore, have deltas ranging from 0 to −1. The short call position lowers the portfolio's overall delta as desired.

3. B is correct. VaR measures the frequency of losses of a given minimum magnitude. Here the VaR indicates that on 5% of trading days, the portfolio will experience a loss of at least \$6.5 million. (Although C may appear to say the same thing as B, it actually implies that the portfolio will experience a loss on 95% of trading days.) The correct interpretation is that returns will be equal to or greater than −\$6.5 million on 95% of trading days; those returns include gains as well as losses.

4. A is correct. The bank policy requires the addition of forward-looking risk assessments, and management is focused on tail risk. Conditional VaR measures tail risk, and stress tests and scenario analysis subject current portfolio holdings to historical or hypothetical stress events.

5. A is correct. VaR measures do *not* capture liquidity risk. "If some assets in a portfolio are relatively illiquid, VaR could be understated, even under normal market conditions. Additionally, liquidity squeezes are frequently associated with tail events and major market downturns, thereby exacerbating the risk."

6. B is correct. VaR is the minimum loss that would be expected a certain percentage of the time over a specified period of time given the assumed market conditions. A 5% VaR is often expressed as its complement—a 95% level of confidence. Therefore, the monthly VaR in Exhibit 5 indicates that \$5.37 million is the minimum loss that would be expected to occur over one month 5% of the time. Alternatively, 95% of the time, a loss of more than \$5.37 million would not be expected.

7. C is correct. Flusk experienced zero daily VaR breaches over the last year yet incurred a substantial loss. A limitation of VaR is its vulnerability to different volatility regimes. A portfolio might remain under its VaR limit every day but lose an amount approaching this limit each day. If market volatility during the last year is lower than in the lookback period, the portfolio could accumulate a substantial loss without technically breaching the VaR constraint.

   A is incorrect because VaR was calculated using historical simulation, so the distribution used was based on actual historical changes in the key risk factors experienced during the lookback period. Thus, the distribution is not characterized using estimates of the mean return, the standard deviation, or the correlations among the risk factors in the portfolio. In contrast, the parametric method of estimating VaR generally assumes that the distribution of returns for the risk factors is normal.

   B is incorrect because a specification with a higher confidence level will produce a higher VaR. If a 99% confidence interval was used to calculate historical VaR, the VaR would be larger (larger expected minimum loss). During the last year, none of Flusk's losses were substantial enough to breach the 5% VaR number (95% confidence interval); therefore, if McKee used a 1% VaR (99% confidence interval), the number of VaR breaches would not change.

8. B is correct. In order to simulate the impact of the latest financial crisis on the current bond portfolio holdings, McKee's valuation model for bonds should use the historical yields of bonds with similar maturity. Historical yields drive the pricing of bonds more than the price history or the current duration. Historical prices for the fixed-income positions currently held in the portfolio may not exist, and even when historical prices do exist, they may not be relevant to the current characteristics (e.g., maturity) of the instrument. Even if the same bonds existed at the time of the latest financial crisis, their durations would change because of the passage of time.

   A is incorrect because using a bond's past price history would mischaracterize the risk of the current portfolio holdings. For this reason, the historical yields are more important in explaining the risks. Historical prices for the fixed-income positions currently held in the portfolio may not exist, and even when historical prices do exist, they may not be relevant to the current characteristics (e.g., maturity) of the instrument.

   C is incorrect because historical bond durations would not capture the current characteristics of the bonds in the portfolio. Duration is a sensitivity measure and is the weighted-average time to maturity of a bond. Even if the same bonds existed at the time of the latest financial crisis, their remaining time to maturity and durations would change because of the passage of time.

9. C is correct. Ming suggested in Analysis 1 to use a historical scenario that measures the hypothetical portfolio return that would result from a repeat of a particular period of financial market history. Historical scenarios are complementary to VaR but are not going to happen in exactly the same way again, and they require additional measures to overcome the shortcomings of the VaR.

10. B is correct. Analysis 2 describes surplus at risk. Surplus at risk is an application of VaR; it estimates how much the assets might underperform the liabilities with a given confidence level, usually over a year.

11. B is correct. Incremental VaR measures the change in a portfolio's VaR as a result of adding or removing a position from the portfolio or if a position size is changed relative to the remaining positions.

12. B is correct. McKee suggests running a stress test using a historical scenario specific to emerging markets that includes an extreme change in credit spreads. Stress tests, which apply extreme negative stress to a particular portfolio exposure, are closely related to scenario risk measures. A scenario risk measure estimates the portfolio return that would result from a hypothetical change in markets (hypothetical scenario) or a repeat of a historical event (historical scenario). When the historical simulation fully revalues securities under rate and price changes that occurred during the scenario period, the results should be highly accurate.

    A is incorrect because marginal VaR measures the change in portfolio VaR given a very small change in a portfolio position (e.g., change in VaR for a $1 or 1% change in the position). Therefore, marginal VaR would not allow McKee to estimate how much the value of the option-embedded bonds would change under an extreme change in credit spreads.

    C is incorrect because sensitivity risk measures use sensitivity exposure measures, such as first-order (delta, duration) and second-order (gamma, convexity) sensitivity, to assess the change in the value of a financial instrument. Although gamma and convexity can be used with delta and duration to estimate the impact of extreme market movements, they are not suited for scenario analysis related to option-embedded bonds.

13. A is correct. VaR has emerged as one of the most popular risk measures because global banking regulators require or encourage the use of it. VaR is also frequently found in annual reports of financial firms and can be used for comparisons.

14. A is correct. VaR is an estimate of the loss that is expected to be exceeded with a given level of probability over a specified time period. The parametric method typically assumes that the return distributions for the risk factors in the portfolio are normal. It then uses the expected return and standard deviation of return for each risk factor and correlations to estimate VaR.

15. B is correct. Value at risk is the minimum loss that would be expected a certain percentage of the time over a certain period of time. Statement 2 implies that there is a 5% chance the portfolio will fall in value by $90,000 (= $6,000,000 × 1.5%) or more in a single day. If VaR is measured on a daily basis and a typical month has 20–22 business days, then 5% of the days equates to about 1 day per month or once in 20 trading days.

16. A is correct. Statement 2 indicates that the Equity Opportunities Fund reported a daily VaR value. One of the limitations of VaR is that it focuses so heavily on left-tail events (the losses) that right-tail events (potential gains) are often ignored.

    B is incorrect because VaR is viewed as forward looking in that it uses the current portfolio holdings and measures its potential loss. The Sharpe ratio represents a backward-looking, return-based measure and is used to assess the skill of the manager.

    C is incorrect because VaR does not provide an accurate risk estimate in either trending or volatile regimes. A portfolio might remain under its VaR limit every day but lose an amount approaching this limit each day. Under such circumstances, the portfolio could accumulate substantial losses without technically breaching the VaR constraint. Also, during periods of low volatility, VaR will appear quite low, underestimating the losses that could occur when the environment returns to a normal level of volatility.

17. C is correct. Measuring VaR at a 5% threshold produces an estimated value at risk of 2.69%.

    From Exhibit 6, the expected annual portfolio return is 14.1% and the standard deviation is 26.3%. Annual values need to be adjusted to get their daily counterparts. Assuming 250 trading days in a year, the expected annual return is adjusted by dividing by 250 and the standard deviation is adjusted by dividing by the square root of 250.

    Thus, the daily expected return is 0.141/250 = 0.000564, and volatility is $0.263/\sqrt{250} = 0.016634$.

    5% daily VaR = $E(R_p) - 1.65\sigma_p = 0.000564 - 1.65(0.016634) = -0.026882$. The portfolio is expected to experience a potential minimum loss in percentage terms of 2.69% on 5% of trading days.

18. C is correct. The change in value of a bond is inversely related to a change in yield. Given a bond priced at B with duration D and yield change of Δy, the rate of return or percentage price change for the bond is approximately given as follows: $\Delta B/B \approx -D\Delta y/(1 + y)$. Under Scenario 3, interest rates decrease by 20 bps. In an environment of decreasing interest rates, the bond with the highest dura-

tion will have the greatest positive return. Bond 3 has a duration of 10.2, which is greater than that of both Bond 1 (duration = 1.3) and Bond 2 (duration = 3.7).

19. C is correct. A traditional asset manager uses *ex post* tracking error when analyzing backward-looking returns. The Diversified Fixed-Income Fund prospectus stipulates a target benchmark deviation of no more than 5 bps. Tracking error is a measure of the degree to which the performance of a given investment deviates from its benchmark.

20. B is correct. Position limits are limits on the market value of any given investment; they are excellent controls on overconcentration. Position limits can be expressed in currency units or as a percentage of net assets. The Alpha Core Equity Fund restricts the exposure of individual securities to 1.75% of the total portfolio.

21. C is correct. The Monte Carlo simulation method can accommodate virtually any distribution, an important factor given the increased frequency of large daily losses. This method can also more easily accommodate the large number of portfolio holdings. The Monte Carlo method allows the user to develop her own forward-looking assumptions about the portfolio's risk and return characteristics, unlike the historical simulation method, which uses the current portfolio and re-prices it using the actual historical changes in the key factors experienced during the lookback period. Given the limited return history for infrastructure investments and Hamilton's expectations for higher-than-normal volatility, the historical simulation method would be a suboptimal choice.

22. C is correct. Conditional VaR is a measure of tail risk that provides an estimate of the average loss that would be incurred if the VaR cutoff is exceeded.

23. C is correct. A hypothetical scenario analysis allows the risk manager to estimate the likely effect of the scenario on a range of portfolio risk factors. A sovereign ratings downgrade would affect Hiram's India equity and corporate bond exposures as well as the government bond exposure. In addition, the assumptions used in constructing the scenario analysis can specifically address the effect of a need to sell large position sizes under decreased liquidity conditions resulting from a ratings downgrade. VaR alone does not accurately reflect the risk of large position sizes, which may be difficult to trade.

24. C is correct. A hypothetical scenario analysis allows Hamilton to estimate the direct effect of a ratings downgrade on the portfolio's government bond holdings and the resulting need to sell a number of the portfolio's holdings because they no longer meet the ratings guidelines. VaR alone does not accurately reflect the risk of large position sizes, which may be difficult to trade. The hypothetical scenario analysis will also highlight the effect of increased economic turmoil on all of the portfolio's exposures, not only the government bond exposures.

25. B is correct. The VaR is derived as follows:

$$\text{VaR} = \{[E(Rp) - 2.33\sigma_p](-1)\}(\text{Portfolio value}),$$

where

$$E(R_p) = \text{Annualized daily return} = (0.00026 \times 250) = 0.065$$

$$250 = \text{Number of trading days annually}$$

$$2.33 = \text{Number of standard deviations to attain 1\% VaR}$$

$$\sigma_p = \text{Annualized standard deviation} = \left(0.00501 \times \sqrt{250}\right) = 0.079215$$

$$\text{Portfolio value} = \text{CAD}260{,}000{,}000$$

$$\text{VaR} = -(0.065 - 0.184571) \times \text{CAD}260{,}000{,}000$$

$$= \text{CAD}31{,}088{,}460.$$

26. B is correct. Given the large fixed-income exposure in the LICIA portfolio, examining the portfolio duration more closely would be prudent. Duration is the primary sensitivity exposure measure for fixed-income investments.

LEARNING MODULE

6

Backtesting and Simulation

by Yin Luo, CPA, PStat, CFA, and Sheng Wang.

Yin Luo, CPA, PStat, CFA, is at Wolfe Research LLC (USA). Sheng Wang is at Wolfe Research LLC (USA).

LEARNING OUTCOMES

Mastery	The candidate should be able to:
☐	describe objectives in backtesting an investment strategy
☐	describe and contrast steps and procedures in backtesting an investment strategy
☐	interpret metrics and visuals reported in a backtest of an investment strategy
☐	identify problems in a backtest of an investment strategy
☐	evaluate and interpret a historical scenario analysis
☐	contrast Monte Carlo and historical simulation approaches
☐	explain inputs and decisions in simulation and interpret a simulation; and
☐	demonstrate the use of sensitivity analysis

INTRODUCTION

1

☐	describe objectives in backtesting an investment strategy

Sarah Koh heads the quantitative research team at Newton Research Pte. SWF Fund, one of Newton's biggest clients, has asked Koh to help develop new investment strategies by rigorously and independently evaluating their risk and return profiles. SWF Fund would like Koh to evaluate the merits of a "value" equity strategy—does owning "cheap" stocks and avoiding (or short-selling) "expensive" stocks add alpha?—as well as two multifactor fundamental strategies that incorporate several other factors besides value. SWF Fund's Investment Committee will use Koh's findings in its decision-making on whether to begin using these strategies.

Koh's work and findings for SWF Fund will be illustrated throughout the reading.

This reading provides an overview of four techniques used to evaluate investment strategies. The first technique, known as **backtesting**, tests a strategy in a historical environment, usually over long periods, answering the question "How would this strategy have performed if it were implemented in the past?" The second technique, **historical scenario analysis**, also known as **historical stress testing**, examines the efficacy of a strategy in discrete historical environments, such as during recessions or periods of high inflation. The third technique, **simulation**, explores how a strategy would perform in a hypothetical environment specified by the user, rather than a historical setting; it is a useful complement to other methods because the past may not recur and only a limited number of all possible future observations for important variables (e.g., interest rates, return correlations, economic growth) is represented in history. Finally, we explore **sensitivity analysis**, which is often combined with simulation to uncover the impact of changing key assumptions.

Increasingly powerful off-the-shelf software has moved these techniques from the realm of specialists to generalists. In a CFA Institute survey of nearly 250 analysts, portfolio managers, and private wealth managers, 50% of respondents reported that they had performed backtesting analysis on an investment strategy in the past 12 months. Although performing these analyses now has fewer technical challenges than before, understanding the steps and procedures, the implicit assumptions, the pitfalls, and the interpretation of results have only increased in importance given the proliferation of these tools. This reading is a starting point on the journey to building this core professional competency.

2 THE OBJECTIVES OF BACKTESTING

Backtesting approximates the real-life investment process by using historical data to assess whether a strategy would have produced desirable results. Although not all strategies that perform well in a backtest will produce excess returns in the future, backtesting can offer investors insight and rigor to the investment process. Conversely, a strategy that does not show efficacy in backtesting could deliver excess returns in the future, but such a strategy is unlikely to be accepted by portfolio managers and investors alike. As a result, backtesting can be employed as a rejection or acceptance criterion for an investment strategy, depending on the investment manager's process.

Backtesting has been widely used in the investment community for many years. Although it fits quantitative and systematic investment styles more naturally, it is also widely used by fundamental managers. Before using a criterion to screen for stocks (such as a valuation metric, for example), a backtest can uncover the historical efficacy of that criterion by determining if its use would have added incremental excess return.

The implicit assumption in backtesting is that the future will at least somewhat resemble history. The reality, however, is more complicated. We attempt to account for the randomness of the future using complementary techniques discussed later in the reading.

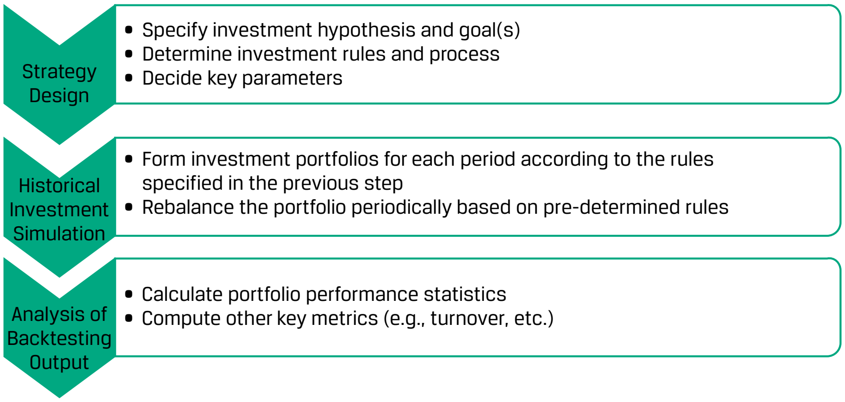
THE BACKTESTING PROCESS

3

- ☐ describe and contrast steps and procedures in backtesting an investment strategy
- ☐ interpret metrics and visuals reported in a backtest of an investment strategy

Backtesting consists of three steps: strategy design, historical investment simulation, and analysis of backtesting output. Exhibit 1 illustrates these steps and component procedures. We will discuss each step and illustrate them with example backtests of two investment strategies.

Exhibit 1: Backtesting Flowchart



Source: Wolfe Research Luo's QES.

Step 1: Strategy Design

The first step is to identify the investment goals and hypothesis. For active strategies, the goal is typically to achieve excess returns over the relevant benchmark or superior risk-adjusted absolute return. An investment hypothesis is a method—a trading rule, security selection criterion, a portfolio, etc.—aimed at achieving the goal.

The next step is to translate the hypothesis into rules and processes and to specify several key parameters, so that the hypothesis can be backtested. The key parameters include the investment universe, specific definition of returns, frequency of portfolio rebalancing, and start and end dates.

Investment Universe

The investment universe refers to all of the securities in which we can potentially invest. Although academic researchers and specialists typically use the union of Compustat/ Worldscope and CRSP,[1] many practitioners use the constituents of well-known broad market indexes as their investment universe. In this reading, unless specified otherwise,

1 CRSP (the Center for Research in Security Prices) provides high-quality data and security returns. The CRSP data series of New York Stock Exchange–listed stocks begins on 31 December 1925.

we use the Russell 3000 Index, S&P/TSX Composite Index, MSCI China A, and S&P Global Broad Market Index (BMI) for the investment universe for equity strategies in the United States, Canada, mainland China, and all other markets, respectively.

Return Definition

As we extend our investment universe from a single country to a global context, multiple complexities arise, such as currency, trading, and regulatory considerations. For example, we need to decide in what currency the return should be computed. The two most frequent choices are either to translate all investment returns into one single currency—typically the home country currency—or to denominate returns in local currencies. The choice of currency in backtesting often depends on whether the portfolio manager hedges their currency exposures. Managers who do not hedge their exchange rate risk often choose to backtest using single-currency-denominated returns.

If the goal of the investment strategy is excess return, a benchmark must also be specified. The benchmark used is often the benchmark for the client mandate or fund for which the investment strategy under study is applicable. The benchmark should relate to the investment universe; for example, the MSCI China A Index is a logical choice for a strategy that uses the constituents of that index as its universe.

Rebalancing Frequency and Transaction Cost

Practitioners often use a monthly frequency for portfolio rebalancing, although higher or lower frequencies are also common. Note that daily or higher frequency rebalancing typically incurs higher transaction costs, and price data will likely be biased by bid–ask spreads, asynchronous trading across different parts of the world, and missing days because of holidays in different countries. Consideration of transaction costs is critical, because many market anomalies simply disappear once they are included. As such, the analyst should explicitly communicate whether transaction costs are included or not in any presentation of the output.

Start and End Date

All else equal, investment managers prefer to backtest investment strategies using as long a history as possible, because a larger sample imparts greater statistical confidence in the results. Conversely, however, because financial data are likely to be non-stationary, performance over a long data history should be supplemented with examinations of discrete regimes within the long history (e.g., periods of high and low inflation, recessions and expansions, etc.) using historical scenario analysis, which we will discuss later in the reading.

EXAMPLE 1

Strategy Design

After an initial conversation with the investment committee at SWF Fund, Sarah Ko notes the following:

Goal:

Superior risk-adjusted absolute return.

Hypothesis:

"Cheap" stocks—those with lower relative valuations—will outperform "expensive" stocks. In other words, exposure to the "value" factor will lead to outperformance.

Koh must now further specify the hypothesis to allow backtesting, as well as define key parameters for the backtest.

The value factor can be described using almost any combination of market price and fundamental performance measures, on a historical (called trailing) or forward-looking basis. Koh selects a simple valuation metric—trailing earnings yield, the inverse of the P/E—to quantify the "cheapness" of a stock. Although P/E is more commonly understood than earnings yield, a serious flaw is that it cannot be computed or logically interpreted if EPS is zero or negative. Earnings yield, on the other hand, can be computed for any stock so long as EPS and price data are available.

$$\text{Trailing earnings yield} = \frac{\text{Trailing 12-month EPS}}{\text{Current share price}}. \quad (1)$$

Specification of Key Parameters:

- Investment universe: Russell 3000 for the US market and S&P Europe BMI for the European market. Total returns will be hedged back into US dollars.
- Start and end date: Because data required for this strategy are widely available, Koh will use a long time period: January 1986–May 2019.
- Rebalancing frequency: monthly, including transaction costs, but returns on a 12-month moving average basis will be computed.

1. Given the backtesting strategy design outlined here, which of the following is a concern about which the investment committee of SWF Fund should be aware?

   A. The strategy assumes that the US dollar will appreciate against the euro.

   B. The historical period of the data includes recessions, currency regime changes, and periods of varying interest rates.

   C. There are serious issues with computing earnings yield for many stocks.

Solution:

B is correct. The portfolio manager is using a long data history that includes regime changes in inflation, currencies, and interest rates, so the data is non-stationary. Consequently, backtesting performance results should be supplemented with examinations of performance during the discrete regimes.

A is incorrect because the analysis makes no assumption about exchange rates. C is incorrect because earnings yield can be computed as long as EPS and price data are available.

2. Which of the following describes the relationship between rebalancing frequency and transaction costs?

   A. Changing the rebalancing frequency from monthly to weekly would likely increase transaction costs.

   B. Changing the rebalancing frequency from monthly to quarterly would likely increase transaction costs.

   C. Rebalancing frequency has no effect on transaction costs.

Solution:

A is correct. Rebalancing frequency refers to how often a portfolio is updated to reflect current data, such as (in this case) changes in earnings yields across the investment universe. Typically, the more frequently rebalancing is done, the more trading is required, which incurs more transaction costs. B is incorrect because it describes a decrease in rebalancing frequency, which would decrease transaction costs. C is incorrect because rebalancing frequency is the primary driver of trading volume, which incurs transaction costs.

3. Which of the following is not a potential concern of using a short time period for a backtest?

   A. The backtest will cover a limited number of business cycle, inflation, and interest rate regimes.

   B. The backtest may not be useful because the findings may apply only under the conditions present in the time frame.

   C. The backtest is likely to cover multiple business cycle, inflation, and interest rate regimes.

Solution:

C is correct. Covering multiple macroeconomic regimes *is not* a concern associated with using a short time period for a backtest, because macroeconomic regimes tend to be multi-year in length. A and B are incorrect because they *are* concerns associated with using a short time period: The backtest may capture only a limited experience, and thus the findings may be relevant for only that experience.

Step 2: Historical Investment Simulation

The next step is constructing the portfolio to be tested and ensuring that it is rebalanced based on the pre-determined frequency.

The portfolio construction process depends primarily on the investment hypothesis under consideration (e.g., whether it is an entire portfolio, a trading strategy, or a modification of an existing strategy), the investment manager's capabilities and style, and the client's investment mandate for which the potential strategy is relevant (e.g., are there geographical limitations? Are there size and liquidity constraints? Can the manager short stocks?). Although our examples use fundamental factor-based, quantitative equity strategies in which we assume stocks can be shorted, backtesting can be applied to any kind of investment strategy.

To simulate rebalancing, analysts typically use **rolling windows**, in which a portfolio or strategy is constituted at the beginning of a period using data from a historical in-sample period, followed by testing on a subsequent, out-of-sample period. The process is repeated as time moves forward. This approach replicates the live investing process, because investment managers adjust their positions as new information arrives. For example, assume we backtest a value strategy by measuring its performance each month from December 2011 to May 2012. The process begins on 30 November 2011 by compiling every stock's trailing 12-month earnings yield using EPS reported in the previous 12 months (i.e., from December 2010 to November 2011, the in-sample months) divided by stock prices as of 30 November 2011. We then execute the investment strategy—for example, buying stocks with high earnings yields and shorting stocks with low earnings yields—as of that date. Then, we record the investment results for the month of December (i.e., the out-of-sample, OOS, month). The process is repeated at the end of each subsequent month, by rebalancing the portfolio with refreshed trailing 12-month earnings yield data and measuring the results over the ensuing (OOS) month. Exhibit 2 illustrates this process is illustrated.

Exhibit 2: Rolling Window Backtesting of the Earnings Yield Factor

	2010:12	2011:01	2011:02	2011:03	2011:04	2011:05	2011:06	2011:07	2011:08	2011:09	2011:10	2011:11	2011:12	2012:01	2012:02	2012:03	2012:04	2012:05
11/30/2011									In-Sample (Last 12M EPS/Price)				OOS					
12/31/2011										In-Sample (Last 12M EPS/Price)				OOS				
1/31/2012											In-Sample (Last 12M EPS/Price)				OOS			
2/29/2012												In-Sample (Last 12M EPS/Price)				OOS		
3/31/2012													In-Sample (Last 12M EPS/Price)				OOS	
4/30/2012														In-Sample (Last 12M EPS/Price)				OOS

Source: Wolfe Research Luo's QES.

Step 3: Analysis of Backtesting Output

The final step in backtesting is generating results for presentation and interpretation. We care about not only the average return of the portfolio but also the risk profile (e.g., volatility and downside risk). Therefore, analysts often use metrics such as the Sharpe ratio, the Sortino ratio, volatility, and **maximum drawdown**. Maximum drawdown is the maximum loss from a peak to a trough for an asset or portfolio.

Beyond these measures, other key performance outputs are visual: for example, time series of returns as well as distributions of returns plotted against a well-known distribution, such as the normal distribution. Visuals are an intuitive way of summarizing many datapoints that often reveal more than a single number summary measure.

It is also useful to examine the backtested cumulative performance of an investment strategy over an extended history. We recommend plotting performance using a logarithmic scale, wherein equal percentage changes are presented as the same vertical distance on the *y*-axis. Using these cumulative performance graphs, one can readily identify downside risk, performance decay, and structural breaks. Structural breaks, or regime changes, are the result of many exogenous factors and are one reason why the past is not always a good guide to the future. The following are examples of structural breaks:

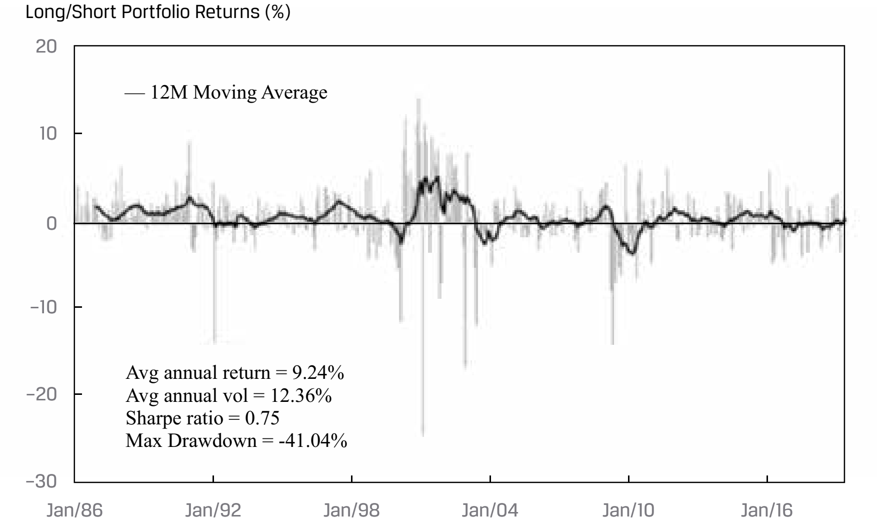
- Depressions and recessions, such as the 2008–09 global financial crisis;
- Geopolitical events, such as changing trade relationships involving countries representing important global equity and bond markets, as well as key countries exiting or entering major trading blocs;
- Major shifts in monetary and fiscal policies, such as the prolonged period of quantitative easing (QE) adopted by major central banks in the aftermath of the 2008 global financial crisis; and
- Major technological changes and advances, such as those that fueled the dot-com bubble and the proliferation of machine learning and artificial intelligence.

We implement our earnings yield-based value strategy as a long–short hedged portfolio, a widely used approach pioneered by Fama and French (1993). In this approach, the analyst sorts the investable stock universe by the relevant metric—trailing earnings yield, in this case—and divides the universe into quantiles (typically into quintiles or deciles) based on those metrics. A long–short hedged portfolio is then formed by going long the top quantile (i.e., the group of stocks with the highest earnings yield) and shorting the bottom quantile (i.e., the group of stocks with the lowest earnings yield). Individual stocks are either equally weighted or market capitalization weighted within each quantile. Although the quantiles may not have equal beta exposure, and they may have exposures to other common factors, the difference in the average earnings yield metric between quantiles is a reasonable and straightforward characterization of underlying performance.

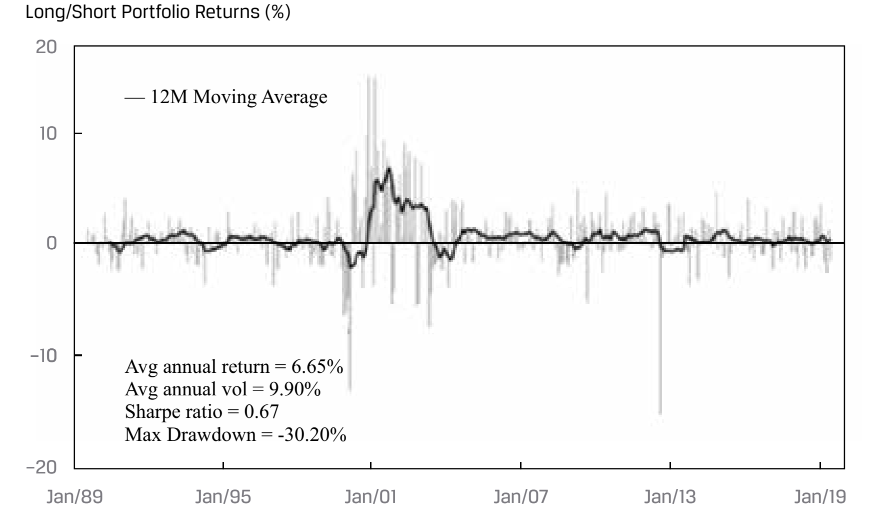
We used quintiles (e.g., top and bottom 20%) and monthly rolling windows and measured the results in several ways for the strategy in the US and European markets, shown in Exhibit 3.

Exhibit 3: Earnings Yield Factor, Long–Short Hedged Quintile Portfolio Returns (January 1986–May 2019)

A. US: Trailing Earnings Yield



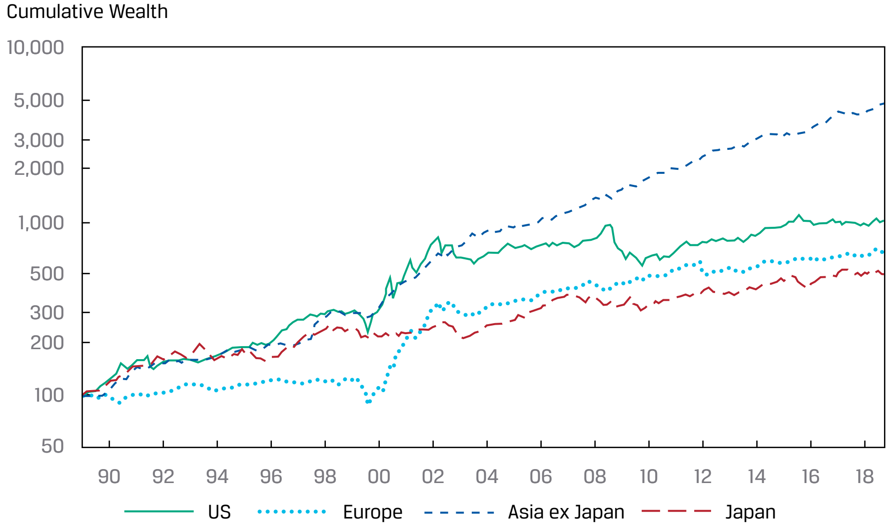
B. Europe: Trailing Earnings Yield



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES.

We also examined the cumulative performance of the strategy in four different stock markets: the United States, Europe, Asia ex-Japan, and Japan, with the results shown in Exhibit 4.

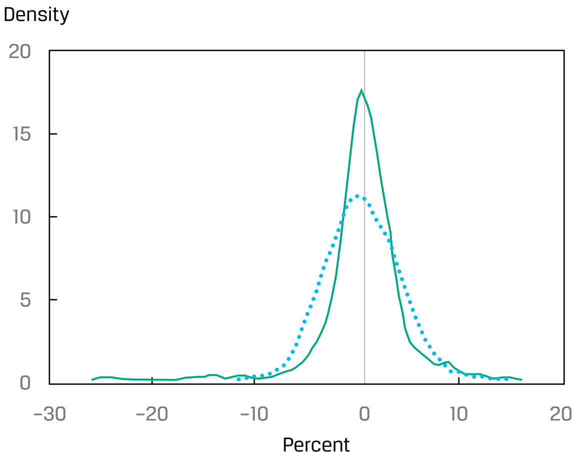
Exhibit 4: Earnings Yield Factor, Long–Short Hedged Quintile Portfolio Returns (January 1986–May 2019) in Several Markets



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, and Wolfe Research Luo's QES.

Finally, we show the distribution of the strategy's returns in the US market against the normal distribution in Exhibit 5.

Exhibit 5: Distribution of Earnings Yield Returns, United States (1986–2019)



Source: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES.

EXAMPLE 2

Historical Investment Simulation and Output Analysis

1. Describe how the backtest performance of value investing, based on the earnings yield factor, in Europe compares with that in the United States over the 1986–2019 period, as shown in Exhibit 3.

Solution:

In the United States, the average annual return from the value investing strategy is about 9.2%, with a Sharpe ratio of 0.75, over the backtesting period (January 1986–May 2019), as seen in Exhibit 3, Panel A. In Europe, the same investment strategy generated a significantly lower (by 250 bps) average annual return, about 6.7%, but with significantly lower volatility (Panel B). Hence, the Sharpe ratio for the European strategy, 0.67, is close to that of the US strategy. In both markets, the maximum drawdown is just over three times the volatility of the strategy. Therefore, as a long-term value strategy, the earnings yield factor offers slightly better performance in the United States than in Europe.

2. Describe the cumulative performance of value investing across the different markets shown in Exhibit 4 and the distributions of returns in the United States from this strategy in Exhibit 5.

Solution:

The value strategy has delivered strong performance over the long run across the several markets, especially in Asia ex-Japan (Exhibit 4). Performance has flattened since 2016, however, in the United States, Europe, and Japan after first leveling off in all geographies except Asia ex-Japan after 2002. Significant drawdowns and potential structural breaks can also be observed in late 1990s (i.e., during the tech bubble) and in March–May 2009 (i.e., the risk rally during the global financial crisis) in most regions.

More problematically, the strategy in the United States seems to suffer from excess kurtosis (i.e., fat tails) and negative skewness (Exhibit 5). The excess kurtosis implies that this strategy is more likely to generate surprises—that is, extreme returns—whereas the negative skewness suggests that those surprises are more likely to be negative (than positive).

4 BACKTESTING MULTIFACTOR MODELS

- ☐ describe and contrast steps and procedures in backtesting an investment strategy
- ☐ interpret metrics and visuals reported in a backtest of an investment strategy

Few investment managers use a single signal, such as earnings yield, in an investment strategy. In practice, most quantitative stock selection models use a multifactor structure, with a linear combination of factors being the dominant framework. Similarly, most fundamental managers use multiple filters in their stock screening tools.

In this section, we introduce two multifactor equity portfolio strategies to more richly illustrate backtesting: a benchmark (BM) factor portfolio, which equally weights multiple fundamental factors, and a **risk parity** (RP) factor portfolio, which weights factors based on equal risk contribution. We chose these two approaches because their weighting schemes—equal weights and equal risk weights, respectively—are objective. We will continue to use these two portfolios throughout the reading to discuss other evaluation techniques.

To backtest these two portfolios, we follow the same three steps described previously: strategy design, historical investment simulation, and output analysis.

Step 1: Strategy Design

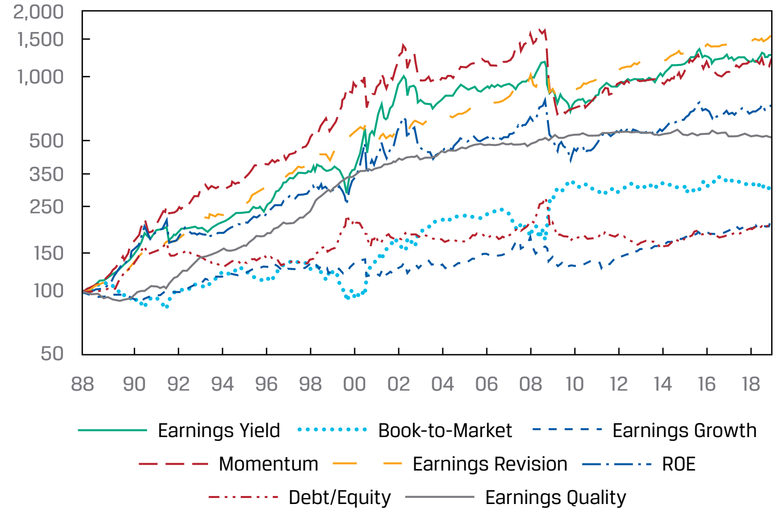
We chose eight fundamental factors from common investment styles:

1. Defensive value: Trailing earnings yield
2. Cyclical value: Book-to-market ratio
3. Growth: Consensus FY1/FY0 EPS growth
4. Price momentum: 12-month total return, excluding the most recent month
5. Analyst sentiment: 3-month EPS revision
6. Profitability: Return on equity (ROE)
7. Leverage: Debt-to-equity ratio
8. Earnings quality: Non-cash earnings (proportion of accruals in earnings)

For each factor, we form a portfolio by buying the top 20% of stocks and shorting the bottom 20% of stocks ranked by the factor. Stocks within both long and short buckets are equally weighted. For illustration purposes, we do not account for transaction costs or other portfolio implementation constraints.

As shown in Exhibit 6 (which uses a logarithm scale on the y-axis), all eight factor portfolios have delivered positive returns over the long term (1988–2019) in the United States. Earnings revision, earnings yield, and price momentum factors produced the highest returns, and the earnings growth and debt/equity factors lagged far behind. The eight factor portfolios appear to share some commonalities. Upon visual inspection, returns seem to fall into three clusters: (1) earnings revision, earnings yield, and price momentum; (2) ROE and earnings quality; and (3) book-to-market ratio, earnings growth, and debt/equity. They also show significant dispersions at times.

Exhibit 6: Cumulative Return of Eight Factor Portfolios, United States (1988–2019)



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES.

For our benchmark portfolio, we combine these eight factor portfolios by equally weighting each one. Researchers have found that such an equally weighted portfolio either outperforms or performs in line with portfolios constructed using more sophisticated optimization techniques (e.g., DeMiguel, Garlappi, and Uppal 2007).

For our risk parity (RP) portfolio, we combine the eight factor portfolios by equally weighting them by their risk contribution. Risk parity is a popular alternative portfolio construction technique that accounts for the volatility of each factor and the correlations of returns among all factors in the portfolio. The objective is for each factor to make an equal (hence "parity") risk contribution to the overall risk of the portfolio.

We backtested our two portfolios in each of the following markets: the United States, Canada, Latin America (LATAM), Europe, the United Kingdom, emerging Europe, Middle East, and Africa (EMEA), Asia ex-Japan, Japan, Australia and New Zealand (ANZ), and mainland China. Both portfolios are rebalanced monthly to maintain equal factor weights or equal factor risk contributions (i.e., risk parity). Although each of the eight underlying factor portfolios is a long–short portfolio, our BM and RP multifactor portfolios are long only, meaning the weights allocated to each factor portfolio are restricted to be non-negative, such that weights for each of the underlying portfolios are all positive and sum to 100%.

Step 2: Historical Investment Simulation

Backtesting a multifactor strategy is similar to the method introduced earlier, but the rolling-window procedure is implemented twice, once at each portfolio "layer."

First, we form eight factor portfolios at each given point in time (i.e., monthly) from 1988 until May 2019 using the rolling-window procedure discussed previously. We then combine these factor portfolios into two multifactor portfolios, each with different weights: equal weighted (BM portfolio) and equal risk weighted (RP portfolio).

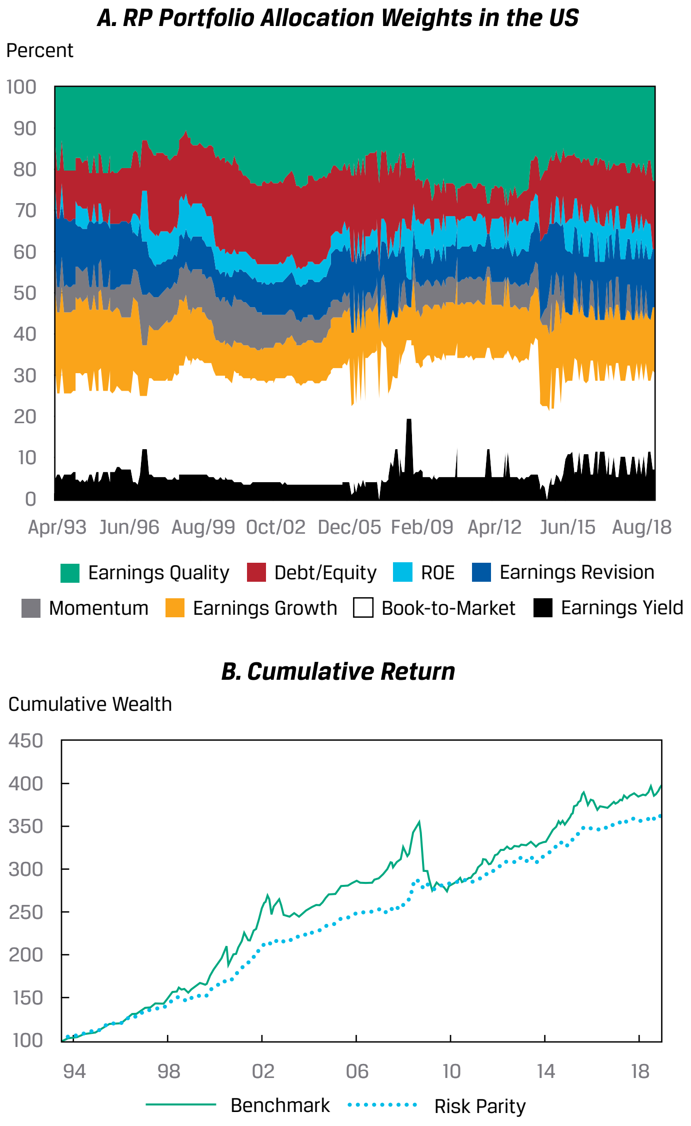
A second rolling-window procedure over the same time span is required to avoid look-ahead bias. At each month end, the previous five years of monthly data are used to estimate the variance–covariance matrix for the eight factor portfolios. This is the most important ingredient to form the RP portfolio. Once the covariance matrix is estimated, we can optimize and compute the weights (i.e., weights for equal risk contribution) for each of the eight factor portfolios and then form the RP portfolio.

Finally, we compute the returns of the two multifactor portfolios (BM and RP) during each "out-of-sample" month from 1988 to May 2019.

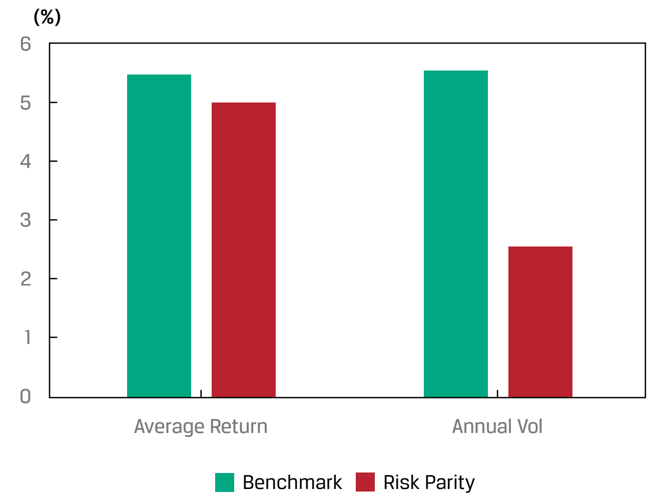
Step 3: Output Analysis

Exhibit 7, Panel A, shows that the weights of the eight factor portfolios in the RP portfolio are relatively stable over time (1993–2019) in the United States, but they are certainly not equal—so we should expect the RP portfolio's risk and return profile to differ from that of the BM portfolio. Notably, book-to-market and earnings quality factor portfolios receive the largest allocations, whereas ROE and price momentum factor portfolios have the lowest weights. Although the RP portfolio appears to deliver a lower cumulative return than does the BM portfolio (Panel B), Panel C shows that the RP portfolio's volatility is less than half the volatility of the BM portfolio. As a result, the RP portfolio's Sharpe ratio is nearly twice that of the BM portfolio (Panel D).

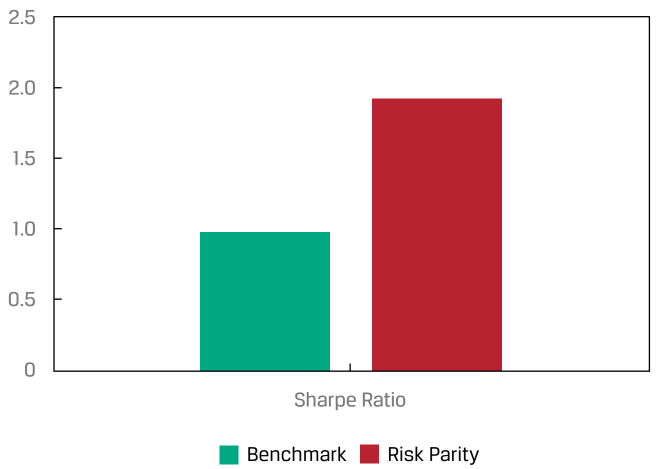
Exhibit 7: Backtesting Multifactor Strategies: Equally Weighted Benchmark Portfolio vs. Risk Parity Weighted Portfolio



C. Average Return and Volatility



D. Sharpe Ratio



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES.

Exhibit 8 presents statistics for the return distributions of the eight factor portfolios and the equally weighted BM and RP weighted multifactor portfolios from 1993 to 2019. Six of the eight factor portfolios have negative skewness (the BM portfolio does as well), and all factors and factor allocation portfolios show excess kurtosis (i.e., kurtosis exceeding 3.0). The downside risk (i.e., minimum monthly return) is clearly greater in magnitude than the maximum upside for most factor strategies. The two factor allocation strategy portfolios—BM and RP—both display moderate mean returns (0.5% and 0.4% per month, respectively) and low standard deviations (1.6% and 0.7% per month, respectively) compared with the eight underlying factor portfolios, highlighting the diversification benefits from factor allocation decisions.

Exhibit 9 compares the various downside risk measures for the eight factor portfolios and the BM and RP portfolios from 1993 to 2019. The three downside risk measures—value at risk (VaR), conditional value at risk (CVaR), and maximum drawdown—suggest that the price momentum factor, followed by the ROE factor, has the largest downside risk. The smallest downside risk is observed for the earnings quality factor. The risk parity portfolio shows considerably less downside risk than any of the eight

Exhibit 8: Monthly Return Distributions: Factor, BM, and RP Portfolios (1993–2019)

	Earnings Yield	Book-to-Market	Earnings Growth	Momentum	Earnings Revision	ROE	Debt/Equity	Earnings Quality	Benchmark	Risk Parity
Mean	0.7%	0.4%	0.2%	0.6%	0.7%	0.5%	0.1%	0.4%	0.5%	0.4%
Median	0.6%	0.1%	0.4%	0.8%	0.8%	0.6%	0.1%	0.4%	0.5%	0.4%
Maximum	14.5%	28.9%	6.2%	11.7%	9.1%	10.8%	11.9%	5.3%	4.3%	3.7%
Minimum	(24.0%)	(12.1%)	(15.8%)	(32.7%)	(18.7%)	(28.0%)	(17.1%)	(2.6%)	(10.9%)	(2.5%)
Std. Dev	3.8%	3.7%	2.1%	4.6%	2.4%	3.9%	2.5%	1.2%	1.6%	0.7%
Skewness	(1.00)	2.82	(2.46%)	(2.36)	(2.39)	(1.92)	(0.58)	0.41	(2.40)	0.51
Kurtosis	11.06	23.61	17.80	16.56	20.76	14.96	11.55	3.87	17.78	5.37

Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES.

underlying factors and the benchmark portfolio. This evidence suggests that the RP strategy benefits greatly from risk diversification, at least in the United States for the period under investigation.

EXAMPLE 3

Risk and Return beyond Normal Distribution

1. Compare return profiles for the BM and RP strategy multifactor portfolios and explain which investment strategy offers the more attractive statistical properties for risk-averse investors (refer to Exhibits 8 and 9).

Solution:

The BM and RP portfolios have nearly the same mean monthly returns, at 0.5% and 0.4%, respectively (Exhibit 8). Although the maximum returns are similar, the RP portfolio has a much smaller minimum return (–2.5%) and a significantly lower standard deviation (0.7%) compared with those of the BM portfolio (–10.9% and 1.6%, respectively). The RP portfolio is also slightly positively skewed (0.51%) and has moderate kurtosis (5.37), in contrast to the negative skew (–2.40%) and high kurtosis (17.78) of the BM portfolio.

The RP portfolio offers similar returns, less downside risk (confirmed by its superior VaR, CVaR, and maximum drawdown results in Exhibit 9), lower volatility, and slightly higher probability of positive returns (i.e., positive skew) compared with the BM portfolio. It is also less fat tailed (i.e., moderate kurtosis, meaning lower probability of extreme negative surprises) than the BM portfolio. Therefore, the RP portfolio has the more attractive distribution properties for risk-averse investors.

EXAMPLE 4

Backtesting the Performance of Factor Allocation Strategies

During the presentation of her backtesting results to SWF's investment committee, Koh is asked the following questions:

1. Regarding rolling-window backtesting, which one of the following statements is *inaccurate*?

   A. The data are divided into just two samples.

   B. Out-of-sample data become part of the next period's in-sample data.

   C. Repeated in-sample training and out-of-sample testing allow managers to adjust security positions on the basis of the arrival over time of new information.

Solution:

A is correct, because the statement is inaccurate. B and C are incorrect, because they accurately describe the rolling-window backtesting technique.

Exhibit 9: Downside Risk Using Monthly Returns: Factor, BM, and RP Portfolios (1993–2019)

	Earnings Yield	Book-to-Market	Earnings Growth	Momentum	Earnings Revision	ROE	Debt/Equity	Earnings Quality	Benchmark	Risk Parity
VaR(95%)	(5.9%)	(0.7%)	(3.9%)	(8.4%)	(3.7%)	(6.8%)	(4.0%)	(1.3%)	(2.6%)	(0.7%)
CVaR(95%)	(14.3%)	(11.1%)	(10.9%)	(22.9%)	(12.8%)	(18.7%)	(8.4%)	(1.7%)	(7.9%)	(0.9%)
Max Drawdown	41.0%	35.3%	27.2%	59.7%	23.9%	47.5%	41.8%	8.3%	22.6%	3.8%

Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES.

2. Which of the following is a drawback of the long–short hedged portfolio approach for implementing factor-based portfolios?

   A. The hedged portfolio is formed by going long the top quantile (with the best factor scores) and shorting the bottom quantile (with the worst factor scores).

   B. Securities must be ranked by the factor being scrutinized and then grouped into quantiles based on their factor scores.

   C. Not every manager can short stocks.

Solution:

C is correct, because it best describes a drawback of the long–short hedged portfolio approach. A and B are incorrect because they describe the approach itself.

3. Which one of the following is *not* a metric or visual used in assessing backtesting of a factor-based investment strategy?

   A. Distribution plots of factor returns

   B. A word cloud of text describing the characteristics of the factor

   C. Maximum drawdown

Solution:

B is correct, because a word cloud is not a visual used in assessing backtesting of a factor-based investment strategy. A and C are incorrect, because they are visuals and metrics, respectively, used to assess backtests of factor-based strategies.

4. Regarding the use of rolling-window backtesting in assessing factor allocation to a risk parity–based strategy, which statement is correct?

   A. The procedure is used once for estimating factor returns over the rolling window.

   B. The procedure is used once for dividing the data into just two samples.

   C. The procedure is used twice—once for estimating factor returns over the rolling window, and a second time for estimating the covariance matrix of factor returns (for deriving risk parity weights) over the rolling window.

Solution:

C is correct, because the procedure must be used a second time for estimating the covariance matrix of factor returns (for deriving risk parity weights) over the rolling window. A is incorrect because the procedure must be done twice: once for estimating factor returns over the rolling window and a second time for estimating the covariance matrix of factor returns (for deriving risk parity weights). B is incorrect because the rolling-window procedure divides the sample into many samples.

5 COMMON PROBLEMS IN BACKTESTING

☐ identify problems in a backtest of an investment strategy

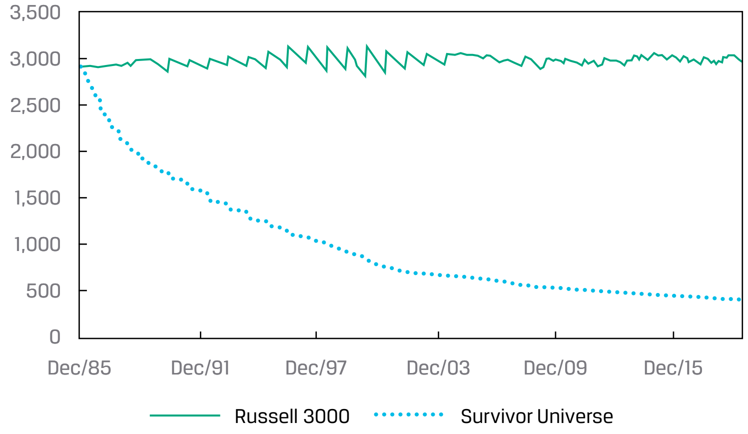
In this section, we discuss some of the most common mistakes investors make when they conduct backtests. Although backtesting is the subject of the discussion, all of these mistakes are relevant to and commonly found in quantitative research generally.

Survivorship Bias

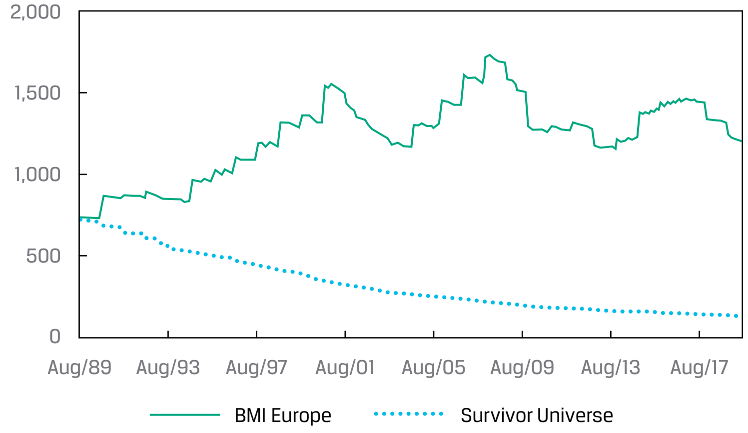
Companies continually appear and disappear from market indexes. New firms appear via IPOs, spin-offs, and outperformance. Companies disappear for many reasons, including privatization, acquisition, bankruptcy, and prolonged under- or outperformance that results in a change in market capitalization from large to mid/small and vice versa. As shown in Panel A of Exhibit 10, fewer than 400 of the constituents of the Russell 3000 Index in 1985 (less than 13%) are still included in the index as of 31 May 2019. Similarly, the S&P BMI Europe Index, which tracks the broad European market, started with about 720 stocks in 1989 and now contains around 1,200 companies. Among the 720 stocks in the index at inception, only 142 (or about 20%) were still in the index as of May 2019 (Panel B of Exhibit 10). Stocks that have remained in the index over time are referred to as “survivors.”

Exhibit 10: Number of Stocks in Index vs. Survivors

A. US (Russell 3000)



B. Europe (S&P BMI)



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES.

Survivorship bias refers to deriving conclusions from data that reflects only those entities that have survived to that date. It is one of the most obvious but, interestingly, also one of the most common mistakes that investors make when conducting backtests. Although the problem is widely covered in the academic literature, relatively few practitioners, whether investing in equities, fixed income, hedge funds, or other asset classes, bother to quantify the implications of survivorship bias in their backtesting.

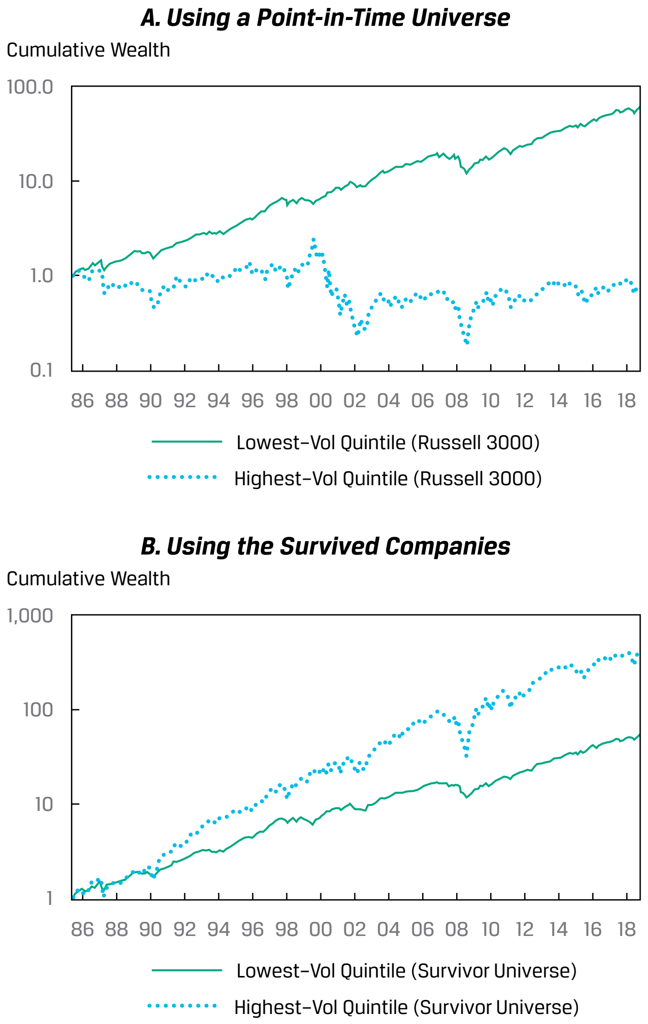
Some investors contend that because you can invest only in companies that exist today, there is nothing wrong with backtesting strategies using only the current index constituents. The problem is, however, that in the past, one could not know which companies would survive in the future, which companies would disappear, and which companies would be created and become successful enough to be added to the index. Moreover, the list of surviving firms is likely biased in one way or another—for example, it could represent primarily multinational firms, or highly innovative firms, or the most successful firms.

Although it is straightforward (but definitely not recommended) to backtest an investment strategy using only the survivors, tracking all companies that have ever existed in a correct point-in-time fashion (i.e., the casualties as well as the survivors) is strongly recommended, especially as such data becomes more available from data vendors. **Point-in-time data** allow analysts to use the most complete data for any given prior time period, thereby enabling the construction (and backtesting) of the most realistic investment strategies.

The difference between backtesting with current index constituents and point-in-time data is illustrated using the low-volatility anomaly, a popular investment strategy that argues that stocks with low volatility tend to outperform high-volatility stocks. A proper backtesting methodology using the point-in-time Russell 3000 universe in Panel A of Exhibit 11 confirms this view; low-volatility stocks have significantly outperformed high-volatility stocks over the three decades up to 2019.

Importantly, however, if we repeat the backtesting exercise using only survivors, then the result is the opposite: high-volatility stocks outperformed low-volatility stocks by about 5.5 times (see Panel B of Exhibit 11). This example underscores the importance of accounting for survivorship bias in backtesting by using point-in-time index constituent stocks and not just the current survivors.

Exhibit 11: Survivorship Bias and the Low-Volatility Anomaly



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES.

Look-Ahead Bias

Another common mistake investors make in backtesting is failing to recognize and account for **look-ahead bias**. This form of bias is created by using information that was unknown or unavailable during the historical periods over which the backtest is conducted. Survivorship bias is actually a type of look-ahead bias, because the question of whether a stock will survive or be added to an index in the future is unknown during the earlier periods over which the backtesting occurs. Look-ahead bias is likely the most common mistake that practitioners make when performing backtesting. It can be overcome by using point-in-time data, which, again, might not be available. Look-ahead bias has several common forms: reporting lags, revisions, and index additions.

The first common form of look-ahead bias derives from reporting lags. For example, in conducting a backtest for year-end 2018, we would not have EPS results for the quarter ending 31 December 2018 for all publicly traded companies until some point around 31 March 2019, although many larger-cap companies might report by

31 January 2019. So, to avoid look-ahead bias, analysts typically compensate by adding several months of reporting lag for every company. This process can also introduce stale information, however. If we continue the example, by 31 January 2019 many larger-cap companies will have already reported earnings, but others, especially mid- and small-cap companies, will not have done so. By using a uniform lag assumption across all companies, the analyst will use stale financial data for some larger-cap companies.

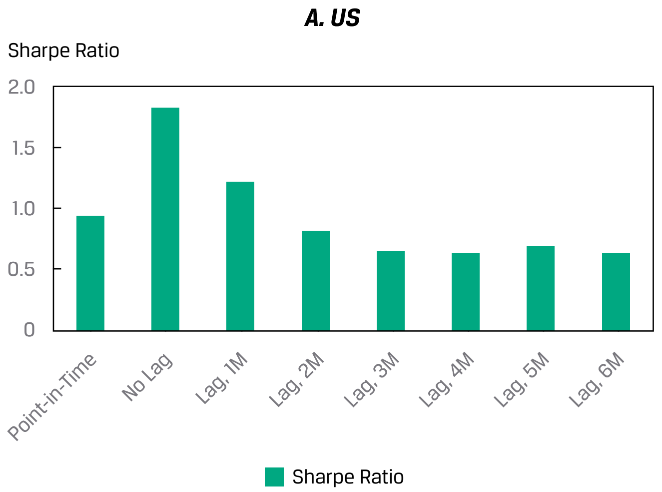
A second problem is data revisions: Macroeconomic data are often revised multiple times, and companies often re-state their financial statements. Many databases keep only the latest numbers, replacing the past figures with the revised ones, although the revised figures were obviously not available at the original release date. By using such revised data, an analyst trying to build realistic investment scenarios going back in time would be using information that was unavailable at that point in time.

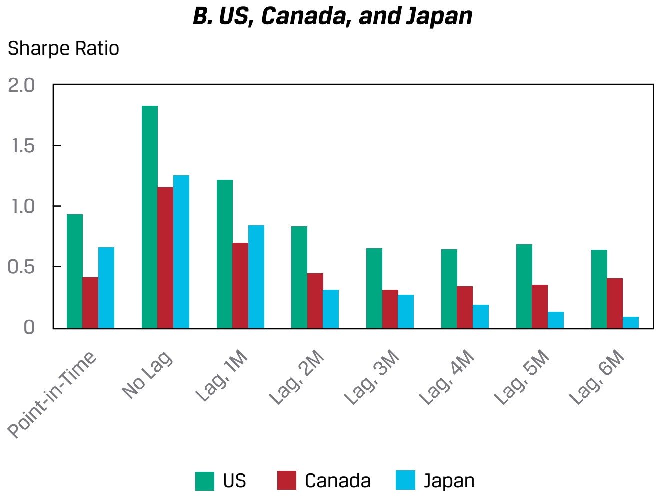
Another form of look-ahead bias arises when data vendors add new companies to their databases. When doing so, they often add several years of historical financial statements into the system. Thus, an analyst backtesting with the current database would be using information on companies that were not actually in the database during the backtesting period. The consequence of this look-ahead bias is often overly optimistic results.

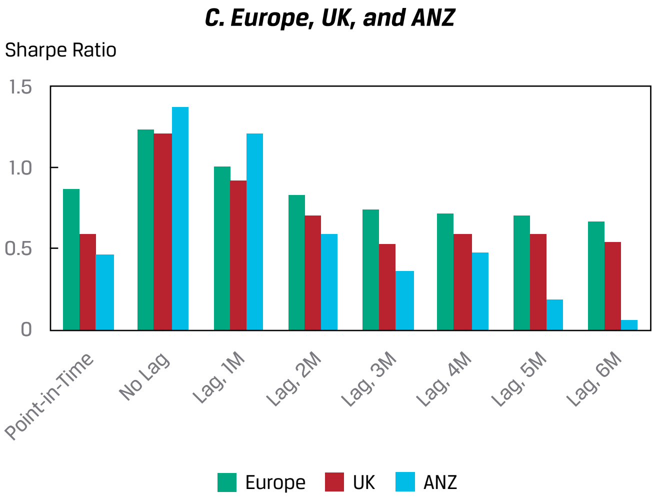
To demonstrate the impact of look-ahead bias and the reporting lag assumption, we conduct monthly backtesting using the earnings yield strategy discussed previously. We compared the backtest results using a proper point-in-time database with the actual EPS data as of each month end, against reporting lag assumptions ranging from zero to six months (a zero lag assumption would suffer from full look-ahead bias).

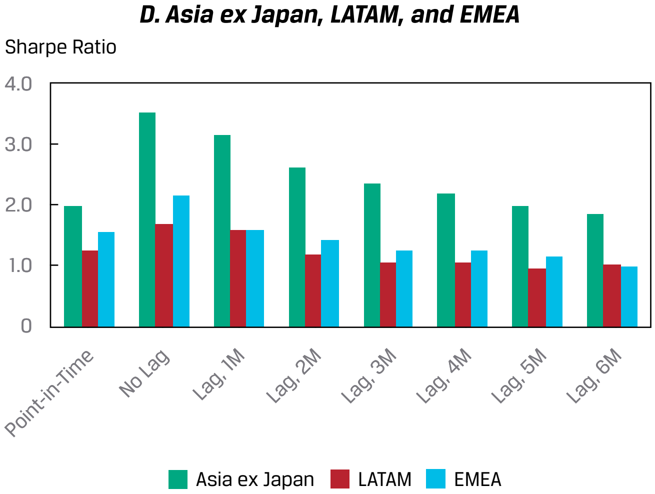
As shown in Exhibit 12, Panel A, it is clear from the backtesting results of the point-in-time scenario against the no-lag scenario that look-ahead bias inflates the performance of our value strategy in the United States by almost 100%. The impact of look-ahead bias is evident in all regions. In the United States, Canada, and Japan (Panel B), it appears that a reporting lag of between one and two months produces backtest results that are consistent with those of the proper point-in-time data. In Europe, the United Kingdom, and ANZ (Panel C), a lag assumption of between two and three months appears appropriate, whereas for Asia ex-Japan, LATAM, and emerging EMEA (Panel D), the point-in-time consistent lag assumption increases to three months. These different lag assumptions reflect the timeliness with which companies in each region report their earnings.

Exhibit 12: Look-Ahead Bias: Impact on Backtesting of Reporting Lag Assumptions (1986–2016)









Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES.

Data Snooping

There is often a temptation to substitute sound portfolio construction by simply backtesting many strategies and picking the best-performing strategy. This bias is called **data snooping**—making an inference after looking at statistical results rather than testing a prior inference. Otherwise known as "p-hacking," data snooping occurs when an analyst selects data or performs analyses until a significant result is found. It can take many forms, including performing interim analyses to decide whether to continue collecting data, using many variables and deciding which to report later; dropping outliers only after performing analyses; and so on. The ultimate results are often false positives.

Data snooping may be mitigated by setting a much higher hurdle than typical—for example, a t-statistic greater than 3.0—for assessing whether a newly discovered factor is indeed adding incremental value (i.e., is statistically significant). Another technique to detect and mitigate data snooping is cross validation, in which the analyst partitions the dataset into training data and testing data (i.e., "validation data") and tests a model built from the training data on the validation data. Rolling window backtesting is a form of cross-validation, albeit in a deterministic and non-random manner, as past periods (i.e., in-sample periods) are used to train a model that is applied to the next (i.e., out-of-sample) period.

EXAMPLE 5

Data Snooping in Investment Management

1. A research analyst has just presented her risk factor–based quantitative/systematic investment model for the UK market to you and several other portfolio managers. She reports the development and backtesting of several different models: The number of factors ranged from 5 to 10, rebalancing periods were monthly and quarterly, and rolling windows were implemented for 5, 15, and 25 years of historical data. She recommends the 10-factor model (with monthly rebalancing) because backtesting of 15 years of data generated the following annualized performance metrics: Sharpe ratio of 3.0 and realized volatility of 1.0%. She also reports a *t*-statistic of 2.5 and a *p*-value of 1.3% for this model of UK market returns, which were the highest and lowest statistics, respectively, of all the models.

   Describe the concerns you should raise around the issue of data snooping for this seemingly very attractive strategy.

Solution:

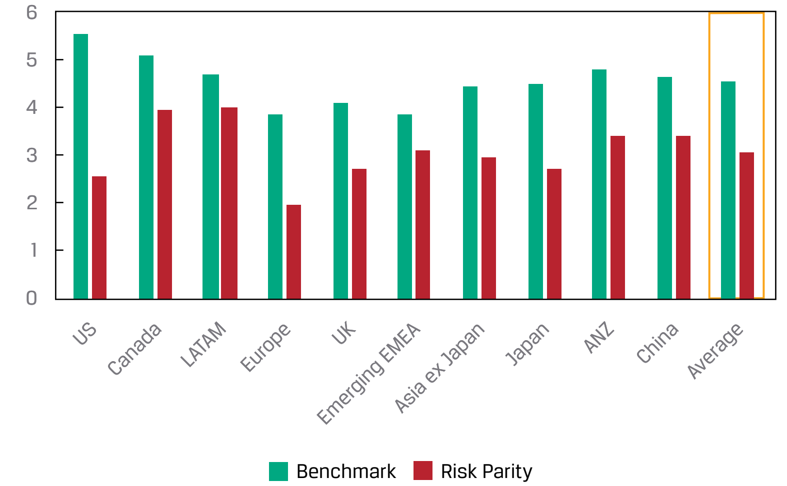
As a portfolio manager, you must be careful in assessing these performance results in light of how the analyst developed and backtested her model. For example, it is critical to know whether backtesting has incorporated transaction costs and trading liquidity. More importantly, however, you need to understand whether data snooping was involved in developing this model/strategy. Given the many variations of models developed and tested by the analyst, it is highly likely that her process suffers from model selection bias. Recommending the model with the highest *t*-statistic and lowest *p*-value also points to data snooping. One way to mitigate the problem is to raise the hurdle for an acceptable model to a *t*-statistic exceeding 3.0 (thereby lowering the *p*-value). The analyst should also consider other techniques that can be used to better understand the true performance of this model/strategy (i.e., cross-validation).

A common way to perform cross validation is to use data from different geographic regions. For example, if the risk parity strategy is developed and tested initially using US equities, the same strategy can be tested in other markets globally to assess whether risk parity is a robust factor allocation strategy.

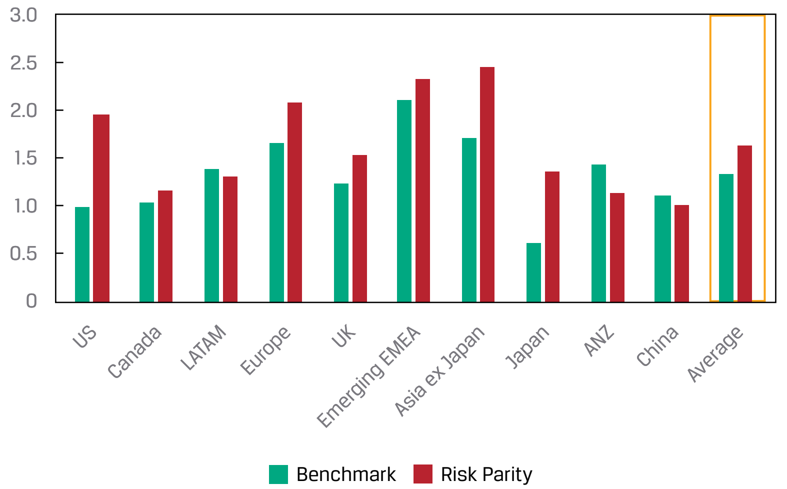
As shown in Exhibit 13, Panel A, as a risk-based factor allocation technique, the RP strategy does indeed deliver a lower realized volatility (i.e., standard deviation of returns) than the benchmark (i.e., equal-weighted factor) strategy in all 10 global markets over 1993–2019. Similarly, the RP portfolios also outperform the BM portfolios in terms of Sharpe ratio (Panel B) in 7 of the 10 global markets.

Exhibit 13: Global Cross-Validation, Equally Weighted Benchmark Portfolio vs. Risk Parity Weighted Portfolio (1993–2019)

A. Realized Volatility



B. Sharpe Ratio



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES.

EXAMPLE 6

Commons Problems in Backtesting

1. An analyst develops an investment strategy by picking the strategy with the highest *t*-statistic and lowest *p*-value after backtesting dozens of different strategies. This approach is an example of which common problem in backtesting?

   A. Reporting lag

   B. Survivorship bias

C. Data snooping

Solution:

C is correct. Data snooping refers to making an inference—such as formulating an investment strategy—*after* looking at statistical results rather than testing a prior inference. A is incorrect because reporting lag refers to the fact that data describing a period is often available only after the period ends and is often subject to revision. B is incorrect because survivorship bias is a form of look-ahead bias in which results are based on a limited, biased sample of subjects (e.g., only surviving companies).

2. Point-in-time data are useful for avoiding the following problems that may affect backtesting *except*:

   A. data snooping.
   B. survivorship bias.
   C. look-ahead bias.

Solution:

A is correct. An analyst can still use a point-in-time dataset to make an inference based on statistical results rather than testing a prior inference. B and C are incorrect, because point-in-time data are useful for avoiding look-ahead bias and survivorship bias (a special case of look-ahead bias). Point-in-time data explicitly corrects for what is not known at a given point in time.

3. The fact that GDP figures for a quarter are not released by government statistical agencies until approximately 30 days after the quarter ends and often undergo several revisions thereafter creates a problem known as:

   A. data snooping.
   B. survivorship bias.
   C. reporting lag.

Solution:

C is correct. Reporting lag refers to the fact that data describing a period is often available only after the period ends and is often subject to revision, which certainly is true of GDP data.

4. Which of the following is an example of cross-validation?

   A. Maximum drawdown
   B. Backtesting with out-of-sample data
   C. Incorporating point-in-time data

Solution:

B is correct. Cross-validation is a technique that involves testing a hypothesis on a different set of data than that which was used to form the inference or initially test the hypothesis. Choice B is the definition of cross-validation.

5. An analyst performed a backtest on an investment strategy in June 2019, selecting the constituents of the Russell 3000 Index as the investment universe, and December 1985 and May 2019 as the start and end dates, respectively. While discussing the results with some colleagues, the analyst

was shown lists of the Russell 3000 Index constituents as of December 2005 and December 1995. She noticed that the lists included only 2,250 and 1,500 companies, respectively, of the Russell 3000 companies at May 2019. The analyst must correct her backtest for which problem?

A. Data snooping

B. Reporting lag

C. Look-ahead bias

Solution:

C is correct. The dataset the analyst uses assumes that the Russell 3000 Index constituents as of May 2019 are the same companies that constituted the index throughout the entire backtesting period. The backtest suffers from look-ahead bias, so conclusions drawn from it will be erroneous because it includes companies that did not exist (or were not index members) over the period starting in December 1985. To correct this problem, the analyst should use a dataset of point-in-time constituents of the Russell 3000 Index.

6 HISTORICAL SCENARIO ANALYSIS

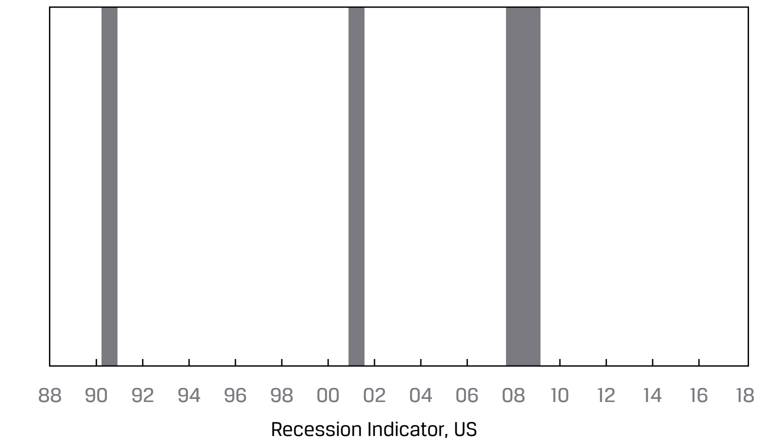
☐ evaluate and interpret a historical scenario analysis

Rather than simply acknowledging or even ignoring structural breaks evident in backtesting results, an analyst should pay careful attention to different structural regimes and impacts to a strategy during regime changes. Historical scenario analysis is a type of backtesting that explores the performance and risk of an investment strategy in different structural regimes and at structural breaks. Two common examples of regime changes are from economic expansions to recessions and from low-volatility to high-volatility environments (and vice versa):

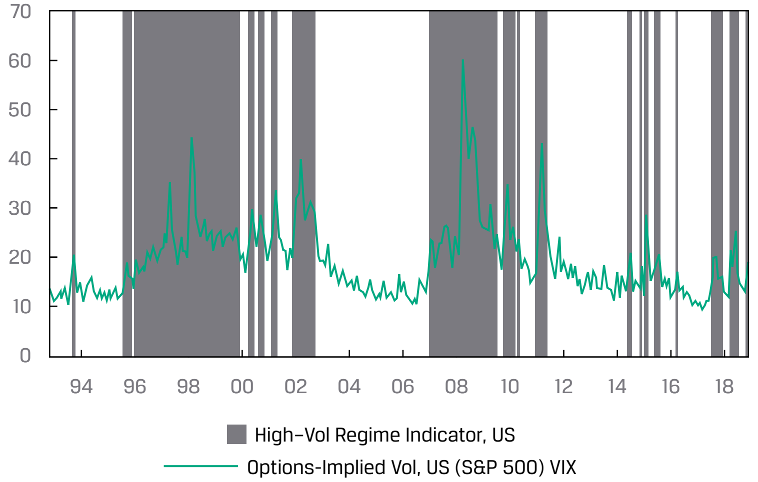
- *Expansions and recessions.* In the United States, since the start of our risk parity allocation strategy in 1993, the National Bureau of Economic Research (NBER) has recognized two official recessions: March 2001–November 2001 and December 2007–June 2009. These recessions are shown in Panel A of Exhibit 14. Although we ignore look-ahead bias in this brief example, it is important to note that business cycle inflection points—the beginning and end of expansions and recessions—are observed only in hindsight. For example, NBER did not identify December 2007 as the beginning of a recession in the United States until December 2008, and it did not identify June 2009 as the end of that recession until September 2010.
- *High- and low-volatility regimes.* The Chicago Board Options Exchange (CBOE) computes the VIX index, which gauges options-implied volatility on the S&P 500 Index. To transform the VIX into a volatility regime indicator, a five-year moving average is computed. Then, the periods when the VIX is above (below) its five-year moving average are defined as high-volatility (low-volatility) regime periods, as shown in Panel B of Exhibit 14 for 1993–2019.

Exhibit 14: Regime Changes

A. Recession Indicator



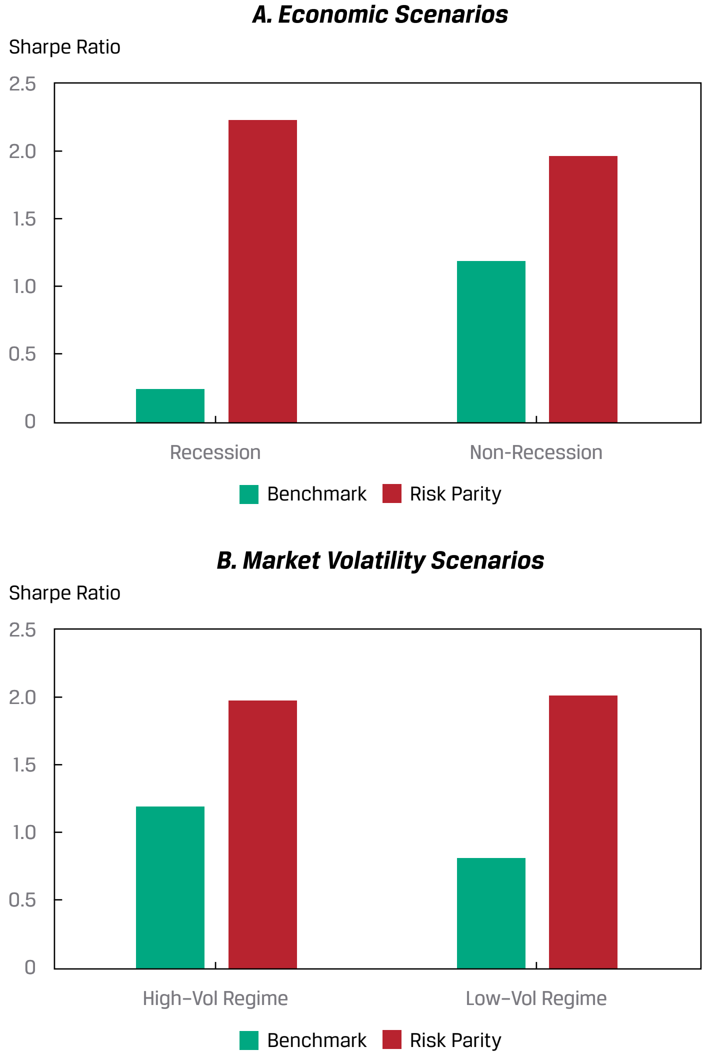
B. VIX: High- vs. Low-Volatility Regimes



Sources: Bloomberg Finance LLP, FTSE Russell, Haver, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES.

We can examine the benchmark and risk parity factor portfolios with respect to these two regimes—recession versus expansion and high volatility versus low volatility. As shown in Panel A of Exhibit 15, in terms of the Sharpe ratio, the RP strategy is quite robust to recession and the BM strategy struggles in recessions. Panel B of Exhibit 15 reveals that the BM strategy's performance is slightly worse in low-volatility regimes than in high-volatility regimes, whereas the RP strategy performs equally well in both volatility environments.

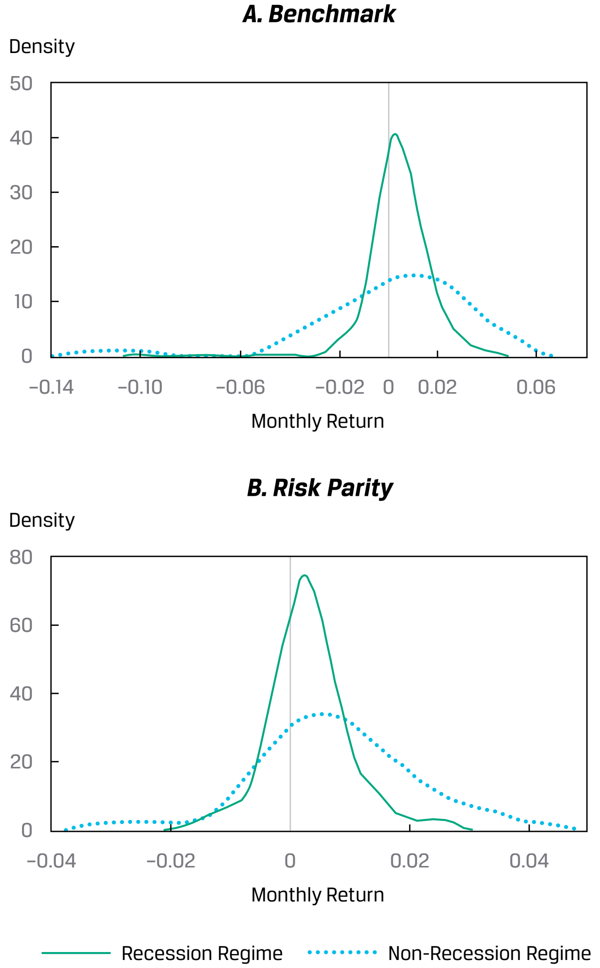
Exhibit 15: Sharpe Ratio for BM and RP Portfolios in Different Macro Scenarios (1993–2019)



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES.

In addition to the Sharpe ratio, a probability density plot can reveal additional information about the sensitivity of the return distributions of these investment strategies—for example, during recession versus non-recession periods. As shown in Exhibit 16, the distribution of returns for both the BM and RP strategies is flatter in a non-recession environment, which implies higher standard deviations during these regimes. The BM strategy suffers from negative skewness and excess kurtosis (i.e., fat tails to the left), regardless of the recession regime, but its average return is clearly lower in a recession environment (Panel A). The RP strategy also has a lower average return in the recession regime (Panel B), but its volatility and kurtosis are both also much lower compared with those of the BM strategy.

Exhibit 16: Distribution of Returns for Factor Allocation Strategies: Recession and Non-Recession Regimes



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES.

SIMULATION ANALYSIS

7

- ☐ contrast Monte Carlo and historical simulation approaches
- ☐ explain inputs and decisions in simulation and interpret a simulation; and

In backtesting, we essentially assume that we can go back in time, apply our investment strategies, rebalance our portfolio(s), and measure performance. This idea is intuitive because it mimics how investing is done in reality—that is, forming our ideas, implementing our strategies, and incorporating new information as it arrives.

Backtesting implicitly assumes that the past is likely to repeat itself, however, and this assumption does not fully account for the dynamic nature of financial markets, which may include extreme upside and downside risks that have never occurred before. We now explore how simulation can provide a more complete picture.

There are two basic types of simulation: historical and Monte Carlo. In **historical simulation**, rather than assuming we implemented a strategy at some past date and collecting results as the strategy runs over time, we instead construct results by selecting returns at random from many different historical periods (windows) without regard to time-ordering. Although this approach does assume, like rolling-window backtesting, that past asset returns provide guidance about future asset returns, it relaxes a key restriction by randomly changing the sequencing of historical periods from which factor returns are drawn. As a result, historical simulation is essentially a non-deterministic rolling-window backtest. Historical simulation is widely used in investment management, particularly by banks for market risk analysis.

The problem with historical time-series data (such as factor returns) is that there is only one set of realized data to draw from—the past happened only one way. A critical assumption behind classical time-series analysis—that the data are stationary—is simply not true of most financial variables. **Monte Carlo simulation** overcomes many of these issues. In Monte Carlo simulation, each key variable is assigned a statistical distribution, and observations are drawn at random from the assigned distribution.

The Monte Carlo approach is popular because it is highly flexible; an array of different distributions can be used across a variety of key variables. Rather than using historical distributions or, for example, the normal distribution (that may only roughly approximate a particular variable's return distribution), the analyst can incorporate non-normality, fat tails, tail dependence, and so on, to model key variables. The downside is that it is complex and computationally intensive.

An important goal of simulation is to verify the investment performance obtained from backtesting by accounting for randomness. Simulation is especially useful in measuring the downside risk of investment strategies.

A properly designed simulation analysis is typically implemented in the following eight steps:

1. Determine what we want to understand: the target variable. This variable is typically the return on an investment strategy or $r_{p,t}$ (the return on portfolio p at time t) and its distribution.
2. Specify key decision variables. Key decision variables are often the returns of each underlying asset, $r_{i,t}$ (the return on asset i at time t), in the overall portfolio and the weight, $\omega_{i,t}$ (the weight of asset i at time t), allocated to each asset in the portfolio. Once we know the returns and weights of all (K) underlying assets, we can readily compute the return of the portfolio as $r_{p,t} = \sum_{i=1}^{K}\left(\omega_{i,t} \times r_{i,t}\right)$. Recall that the weight of each underlying asset is determined by the investment strategy being tested.
3. Specify the number of trials (N) to run. In practice, researchers typically choose between 1,000 and 10,000 simulation runs. The greater the number of trials, the more stable the predictions of performance and variance of performance. In theory, determining the optimal number of iterations is a complex topic (for an example, see Ritter, Schoelles, Quigley, and Klein 2011).
4. Define the distributional properties of the key decision variables. At this point, historical and Monte Carlo simulations diverge. In historical simulation, we draw from historical data. Conversely, in Monte Carlo simulation, we must specify a statistical distribution for each key decision variable. Although it is up to the user, the choice of distribution should be guided by

how well it has described historical observations. It might be appropriate to specify different functions (e.g., normal, lognormal, binomial) for different variables to account for the impact of correlations and tail dependence.

5. Use a random number generator to draw N random numbers for each key decision variable.
6. For each set of simulated key decision variables, compute the value of the target variable. The value of the target variable is then saved for later analysis.
7. Repeat the same processes from Steps 5 and 6 until completing the desired number of trials (N).
8. Now we have a set of N values of the target variable. In this context, it is N returns of the investment strategy. The analyst can now calculate the typical metrics, such as mean return, volatility, Sharpe ratio, and the various downside risk metrics. For simulations, analysts typically use CVaR and maximum drawdown to characterize downside risk.

Historical Simulation

Although backtesting and historical simulation rely on history to understand the future, they are different in that rolling-window backtesting is deterministic, whereas historical simulation incorporates randomness by randomly drawing returns from historical data rather than following each period chronologically.

First, a decision must be made about whether to sample from the historical returns with replacement or without replacement. Random sampling with replacement, also known as **bootstrapping**, is often used in investment research because the number of simulations needed is often larger than the size of the historical dataset.

Using the factor allocation strategies (BM and RP) for the eight factor portfolios as an example, we can perform a historical simulation as follows:

1. The target variables are the returns for the BM and RP multifactor portfolios.
2. The key decision variables are the returns of the eight underlying factor-based portfolios (the weights allocated to the eight factors are already known).
3. The simulation will be performed for N = 1,000 trials.
4. The historical simulation will be implemented using bootstrapped sampling. In this case, we will randomly draw a number from a uniform distribution (so there is equal probability of being selected) between 0 and 1.[2] Once a random number is generated, it is assigned to a specific historical month. Note that we have a total of 374 months of historical factor return data (April 1988–May 2019). We assign random numbers to specific months by dividing the span of the uniform distribution by the number of months (1.0/374 = 0.00267). Therefore, if the random number is between 0 and 0.00267, the first month is selected. Similarly, if the random number generator draws a number between 0.00267 and 0.00535 (= 2 × 0.00267), the second month is chosen, and so on.
5. The random number generator will then randomly draw 1,000 numbers from the uniform distribution between 0 and 1, and, as mentioned, sampling of the historical return data is with replacement. For example, as

2 Technically, the random number generator will draw a random number that equals or is greater than 0 but is less than 1.

shown in Exhibit 17, the first five numbers generated are 0.59163, 0.32185, 0.76485, 0.89474, and 0.45431, which are then mapped to Months 222 (September 2006), 121 (April 1998), 287 (February 2012), 335 (February 2016), and 170 (May 2002), respectively. To be clear, months are mapped by dividing the random number by 0.00267, so Month 222 is determined as 0.59163/0.00267, Month 121 is 0.32185/0.00267, and so on.

Exhibit 17: Factor Returns for the First Five Randomly Selected Months

Simulation #	Month	Random #	Month #	Earnings Yield	Book-to-Market	Earnings Growth	Momentum
1	9/30/2006	0.59163	222	2.5%	0.3%	(0.8%)	(0.0%)
2	4/30/1998	0.32185	121	0.1%	0.8%	(0.2%)	(0.5%)
3	2/29/2012	0.76485	287	(1.9%)	0.5%	1.7%	1.8%
4	2/29/2016	0.89474	335	2.5%	2.4%	(0.4%)	(1.5%)
5	5/31/2002	0.45431	170	6.3%	(3.3%)	1.8%	2.4%

Simulation #	Month	Random #	Month #	Earnings Revision	ROE	Debt/Equity	Earnings Quality
1	9/30/2006	0.59163	222	(0.8%)	2.5%	0.5%	(0.5%)
2	4/30/1998	0.32185	121	(0.1%)	(0.1%)	0.3%	1.6%
3	2/29/2012	0.76485	287	1.8%	(0.5%)	(2.1%)	(0.8%)
4	2/29/2016	0.89474	335	(1.5%)	1.2%	(1.2%)	1.3%
5	5/31/2002	0.45431	170	2.4%	6.4%	(0.7%)	(1.2%)

Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES.

6. Once a given month is selected, the returns of the corresponding eight factor portfolios represent one possible set of outcomes that we use to compute the values of our target variables—the returns of the BM and RP portfolios—using the prespecified factor weights. For example, the first trial picks the month of September 2006. The return of the benchmark portfolio is the equally weighted average of the eight factor returns, or 0.46% (= 0.125 × 2.5% + 0.125 × 0.3% + 0.125 × −0.8% + 0.125 × 0.0% + 0.125 × −0.8% + 0.125 × 2.5% + 0.125 × 0.5% + 0.125 × −0.5%).

To compute the return on the risk parity portfolio, we use the weights allocated to each of the eight factors for the final month (May 2019). As shown in Exhibit 18, for the first trial, September 2006, the weighted average return of the risk parity portfolio is 0.17%. It should be clear that each trial in the historical simulation assumes the simulated returns of the eight factors follow the same patterns observed in the sampled month—in this case, September 2006.

Exhibit 18: How to Compute the Return of the Risk Parity Portfolio, Historical Simulation

Asset (Factor)	September 2006 Return	May 2019 Weight	Weighted Return
Earnings yield	2.5%	6.0%	0.2%
Book-to-market	0.3%	30.3%	0.1%
Earnings growth	(0.8%)	11.7%	(0.1%)
Momentum	(0.0%)	5.2%	(0.0%)
Earnings revision	(0.8%)	10.4%	(0.1%)
ROE	2.5%	6.3%	0.2%
Debt/equity	0.5%	9.6%	0.0%
Earnings quality	(0.5%)	20.4%	(0.1%)
Risk Parity Portfolio			**0.17%**

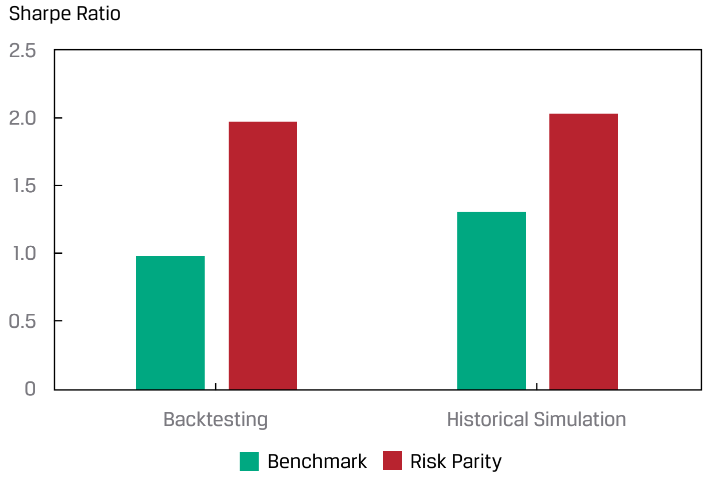
Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES.

7. The same simulation process (from Steps 5 to 6) is repeated for all 1,000 trials, generating a collection of 1,000 simulated returns for the benchmark and risk parity portfolios.
8. Finally, equipped with these 1,000 return scenarios, we can calculate performance metrics of interest (Sharpe ratio, CVaR, etc.) and plot the distributions of the *simulated* benchmark and risk parity portfolio returns.

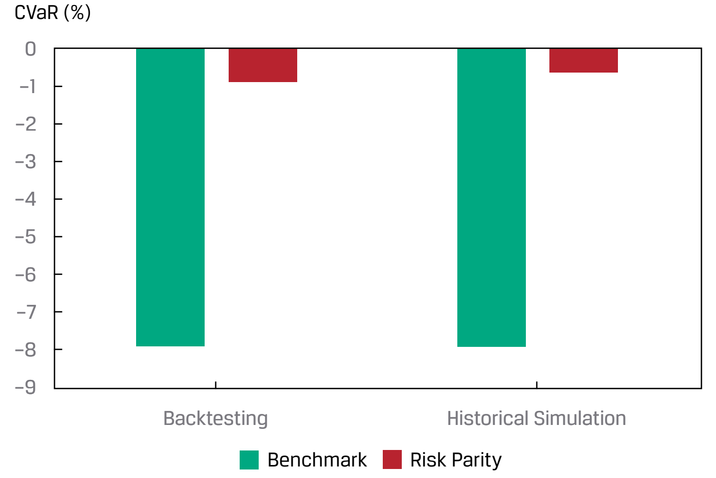
As shown in Panel A of Exhibit 19, the results of the historical simulation (over the 1,000 iterations) suggest that the Sharpe ratios of the BM and RP strategies are largely in line with the rolling-window backtesting method demonstrated previously. In particular, the RP portfolio outperforms the BM portfolio in terms of Sharpe ratio according to both methodologies. Similarly, as shown in Panel B, both methodologies indicate that the RP portfolio carries substantially less downside risk, measured by CVaR, than the BM portfolio.

Exhibit 19: Comparing Historical Simulation with Backtesting

A. Sharpe Ratio



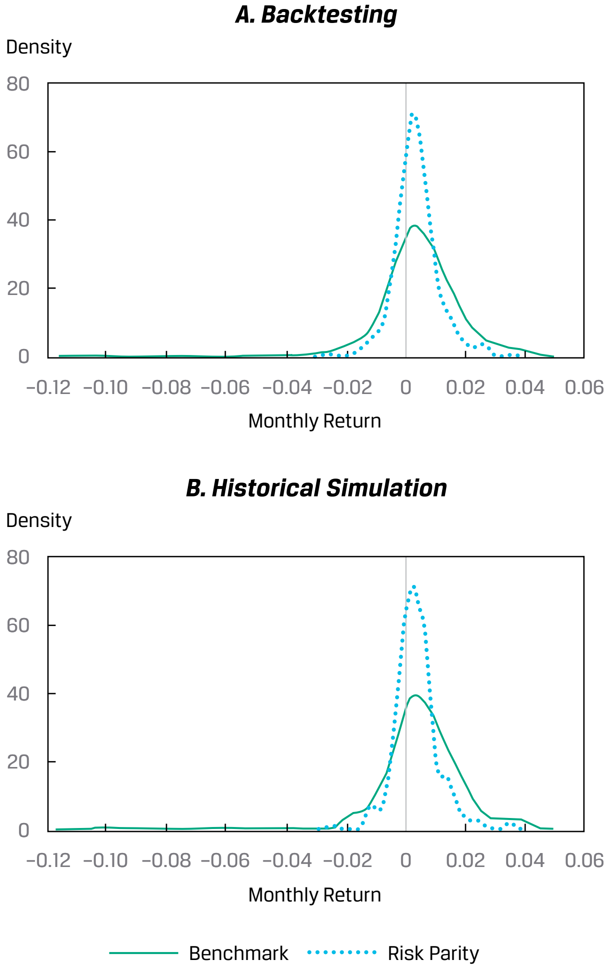
B. Conditonal Value-at-Risk



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES.

In addition to capturing downside risk with a single number (e.g., CVaR), we can also plot the estimated probability distribution of returns for our two investment strategies. Panel A of Exhibit 20 plots the estimated probability distribution of returns for the BM and RP portfolios using backtested returns, whereas Panel B shows the estimated return distribution plots using the historical simulated returns. We can observe a broadly similar pattern between them. Both the backtesting and historical simulation approaches suggest that the RP portfolio returns are less volatile and more skewed to the right with lower downside risk (i.e., lower standard deviation and thinner tails) than the BM portfolio returns.

Exhibit 20: Estimated Distribution Plots: Backtesting and Historical Simulation



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES.

Monte Carlo Simulation

An important issue with historical simulation is that the data are limited to historical observations, which may not represent the future. This deficiency can be addressed with Monte Carlo simulation, which follows similar steps as historical simulation but with a few key differences.

First, we need to specify a functional form for each key decision variable. Exploratory data analysis—focusing on moments (i.e., mean, standard deviation, skewness, kurtosis) and tail dependence—is often crucial here. The usefulness of the Monte Carlo simulation technique critically depends on whether the functional form of the statistical distribution that we specify accurately reflects the true distribution of the underlying data. Because the data's true distribution is unknown, we need to be aware of the fact that our model, like all models, only provides guidance and will not be perfect.

Regression and distribution-fitting techniques are used to estimate the parameters (i.e., mean, standard deviation, skewness, kurtosis) underlying the statistical distributions of the key decision variables. This step is typically called model calibration. Although it may sound difficult, R, Python, Matlab, and similar tools can readily perform this task with a few lines of code (for example, see the fMultivar package in R).

Before finalizing our choice of the functional form of the statistical distribution, we need to account for the following considerations:

- The distribution should reasonably describe the key empirical patterns of the underlying data. For example, asset returns roughly follow a bell curve pattern; therefore, the normal distribution and Student's *t*-distribution are often used as first-cut approximations.
- It is equally critical to account for the correlations between multiple key decision variables. In the case of asset or factor allocation strategies, as shown previously, the returns from multiple factors are clearly correlated; therefore, we need to specify a multivariate distribution rather than modeling each factor or asset on a standalone basis.
- The complexity of the functional form and number of parameters that determine the functional form are important. We can specify a highly complex model with many parameters (all of which need to be estimated/calibrated from historical data) that describe the empirical properties of the data well. Given limited historical data, however, we may be unable to estimate all the underlying parameters with sufficient precision. Such models tend to have low specification errors, but they suffer from large estimation errors. At the other extreme, overly simplistic models require fewer parameters (therefore, they might have low estimation errors), but they may not fit the data well (because they are mis-specified). You should recognize this phenomenon as the bias–variance trade-off, introduced in earlier readings on machine learning and big data projects.

For simulation of asset or factor allocation strategies, the distribution of asset or factor returns is typically modeled as a multivariate normal distribution—as a first-cut approximation—which captures some of the key properties of the underlying data reasonably well. More importantly, a multivariate normal distribution can be fully specified with only a few key parameters—the mean, the standard deviation, and the covariance matrix. For K assets, we need to estimate K mean returns, K standard deviations, and $[K \times (K—1)]/2$ correlations.

We have to be aware, however, that the multivariate normal distribution does not fully account for the empirical characteristics of (negative) skewness, excess kurtosis, and tail dependence apparent in the data. We will address these non-normal distribution properties shortly, when we cover sensitivity analysis.

Continuing with the same BM and RP strategies, the Monte Carlo simulation is performed as follows:

1. Our target variables are the returns for the BM and RP multifactor portfolios.
2. The key decision variables are the returns of the eight underlying factor-based portfolios.
3. We will perform the simulation using 1,000 trials.
4. We choose the multivariate normal distribution as our initial functional form. We calibrate the model—calculate the eight factor portfolio mean returns, the eight standard deviations, and the 28 elements of the covariance matrix—using the 374 months of historical factor return data (April 1988–May 2019).

5. The calibrated multivariate normal distribution is then used to simulate the future factor returns. The process by which this simulation occurs in the context of a multivariate normal distribution of eight random variables, corresponding to our eight factor portfolios, is complex. Suffice it to say, in this case, eight randomly generated numbers from the uniform distribution are mapped onto a point on the joint cumulative probability distribution function, and this point jointly determines the values of the eight factor returns in this trial.

   Exhibit 21 shows the first five sets of Monte Carlo simulated returns for the eight underlying factor-based portfolios.

Exhibit 21: Monte Carlo Simulation: First Five Simulations of Factor Returns Using a Multivariate Normal Distribution

Simulation #	Earnings Yield	Book-to-Market	Earnings Growth	Momentum	Earnings Revision	ROE	Debt/ Equity	Earnings Quality
1	(3.2%)	(3.1%)	(0.2%)	0.7%	2.3%	(3.3%)	(1.7%)	1.9%
2	(0.0%)	3.5%	0.9%	(0.4%)	0.9%	(2.4%)	(3.5%)	(0.2%)
3	0.7%	(1.8%)	2.9%	3.8%	2.5%	1.3%	(0.8%)	(0.0%)
4	9.7%	(0.5%)	1.2%	3.8%	(0.9%)	7.6%	(3.7%)	1.6%
5	1.7%	0.2%	2.9%	(0.2%)	3.0%	0.2%	(0.9%)	0.2%

Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES.

6. Once the returns of the eight factor portfolios are simulated, we can compute the values of our target variables—the returns of the BM and RP portfolios. For example, for the first simulated set of returns, the benchmark portfolio (with equally weighted factor returns) delivers a monthly return of –0.83% (= 0.125 × –3.2% + 0.125 × –3.1% + 0.125 × –0.2% + 0.125 × 0.7% + 0.125 × 2.3% + 0.125 × –3.3% + 0.125 × –1.7% + 0.125 × 1.9%).

   Similarly, using the RP allocation factor weights for the final month, May 2019 (see Exhibit 18), the simulated risk parity portfolio return is –0.86% (= 0.06 × –3.2% + 0.303 × –3.1% + 0.117 × –0.2% + 0.052 × 0.7% + 0.104 × 2.3% + 0.063 × –3.3% + 0.096 × –1.7% + 0.204 × 1.9%).

7. Next, we repeat Steps 5 and 6 for all 1,000 trials to generate a collection of 1,000 returns for the benchmark and risk parity portfolios.

8. Finally, we assess the performance and risk profiles of our two investment strategies from the 1,000 simulated returns.

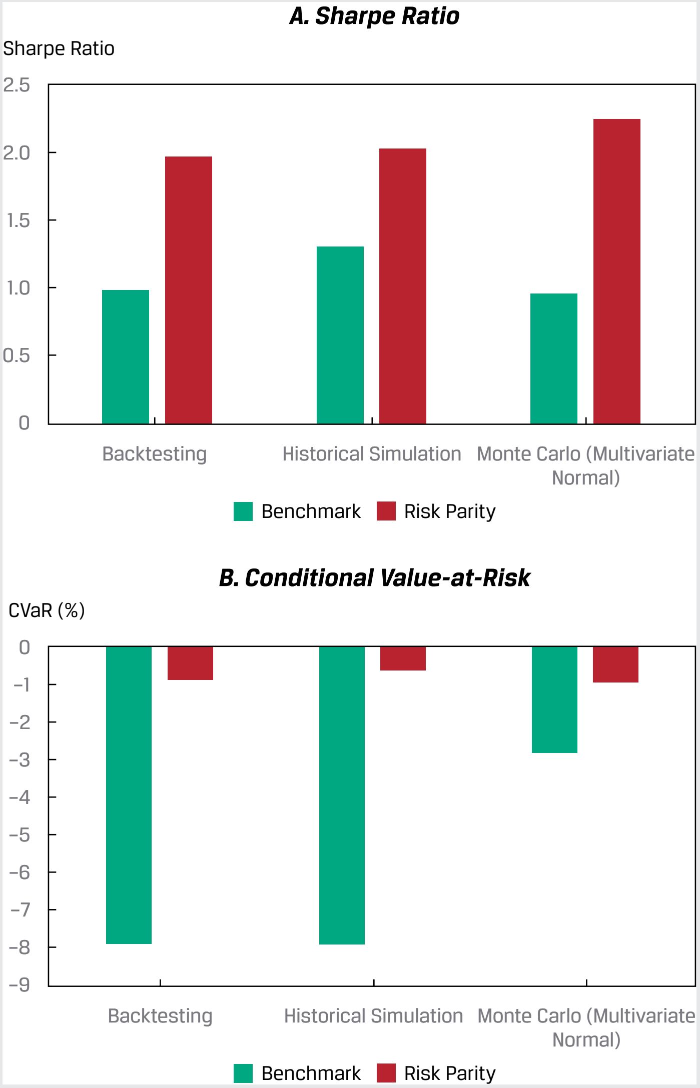
EXAMPLE 7

How to Interpret Results from Historical and Monte Carlo Simulations

1. Exhibit 22 shows the Sharpe ratios (Panel A) and downside risk measures, CVaRs (Panel B), for the returns of the benchmark and risk parity portfolios based on rolling-window backtesting, historical simulation, and Monte Carlo simulation of the returns on the eight underlying factor portfolios.

   Discuss similarities and differences among the three approaches for simulated performance of the benchmark and risk parity portfolios.

Exhibit 22: Comparing Backtesting, Historical Simulation, and Monte Carlo Simulation-Based Performance for the BM and RP Portfolios



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES.

Solution:

Note that the backtesting approach provides realistic performance metrics assuming investors have been following the same trading rules throughout the past periods under investigation. The two simulation analyses are complementary to backtesting and deliver additional insights. In particular, they account for the random nature of investment data in different ways. Historical simulation randomly samples (with replacement) from the past record of asset returns, in a manner that each set of past monthly returns is equally likely to be selected. Monte Carlo simulation randomly samples from an assumed multivariate joint probability distribution (e.g., normal or another type of distribution), in a manner that the past record of asset returns is used to calibrate the parameters of the multivariate distribution. Therefore, these simulation methods are used to independently verify the results from the rolling-window backtesting.

As shown in Panel A of Exhibit 22, the Sharpe ratio appears relatively insensitive to the simulation and backtesting methods used, with the RP strategy outperforming the BM strategy by nearly the same margin for each method. In contrast, CVaR seems to be sensitive to how randomness is treated. In particular, the Monte Carlo simulation appears to understate the downside risk of the BM strategy compared with both rolling-window backtesting and historical simulation methods (Panel B). Because the factor returns are negatively skewed with fat tails (i.e., excess kurtosis), the multivariate normal distribution assumption is likely to be underestimating the true downside risk of the BM strategy. This underestimation of risk appears only for the BM strategy because factor risks and correlations are not properly accounted for in the naive (equal) weighting scheme. Conversely, in this case, the risk parity strategy is robust to a non-normal factor return distribution, resulting in a portfolio with considerably lower downside risk.

SENSITIVITY ANALYSIS 8

- [ ] demonstrate the use of sensitivity analysis

In addition to simulation, sensitivity analysis—a technique for exploring how a target variable is affected by changes in input variables (e.g., the distribution of asset or factor returns)—can be implemented to help managers further understand the potential risks and returns of their investment strategies.

The Monte Carlo simulation just described fits a multivariate normal distribution to the factor returns—a sensible first approximation because it requires relatively few parameters to be estimated from historical data. Despite the simplicity and wide adoption in practice, the multivariate normal distribution assumption fails to account for various empirical properties in the factor return distributions, including negative skewness and fat tails. Because the value of the simulation results depends crucially on whether the selected functional form is a reasonable proxy for the true distribution, we should conduct a sensitivity analysis by fitting our factor return data to a different distribution and repeating the Monte Carlo simulation accordingly. One alternative to test is a multivariate skewed Student's *t*-distribution.

The Student's *t*-distribution is a natural extension of the multivariate normal distribution, because it has the ability to account for the skewness and the excess kurtosis often observed in factor and asset return data. It is mathematically more complex, however, and requires estimating a larger number of parameters than a normal distribution.

With the goal of determining the sensitivity of our target variables (the returns of the benchmark and the risk parity portfolios) to the new factor return distribution assumption, the procedure for the new Monte Carlo simulation process is almost identical to the one performed previously. The only two exceptions are Steps 4 and 5. In Step 4, instead of fitting the data to a multivariate normal distribution, we calibrate our model to a multivariate skewed *t*-distribution. In Step 5, we simulate 1,000 sets of factor returns from this new distribution function. Then, as before, we assess the performance and risk profiles of our investment strategies from the 1,000 simulated returns.

Exhibit 23 shows the first five sets of simulated factor returns from this new model. As previously, we compute the values of our target variables for each set of simulated factor returns and then assess their performance and risk characteristics. For the first set of factor returns, the equal-weighted (i.e., 0.125 for each factor) benchmark portfolio achieves a simulated monthly return of 1.21%, and the risk parity portfolio (using May 2019 factor weights in Exhibit 18) delivers a simulated return of 0.75%.

Exhibit 23: First Five Simulations of Factor Returns Using Multivariate Skewed *t*-Distribution

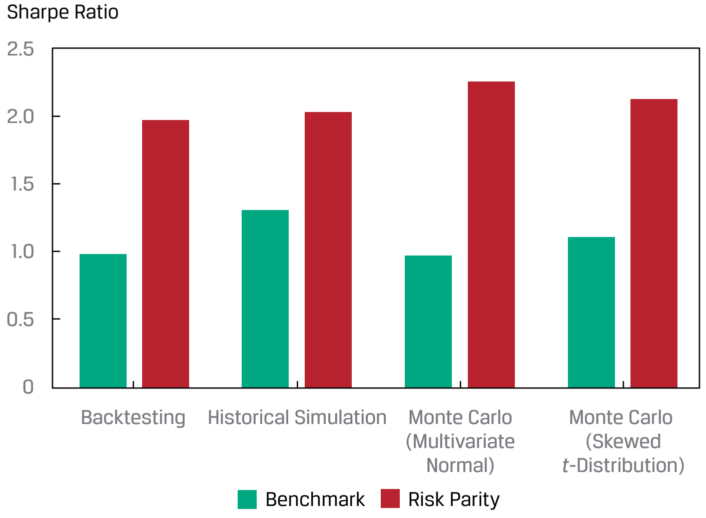
Simulation #	Earnings Yield	Book-to-Market	Earnings Growth	Momentum	Earnings Revision	ROE	Debt/ Equity	Earnings Quality
1	2.0%	0.3%	1.7%	3.1%	2.0%	0.9%	0.2%	(0.5%)
2	1.8%	(1.4%)	0.2%	4.9%	1.8%	2.7%	0.4%	(0.1%)
3	(0.6%)	0.2%	(1.0%)	(0.1%)	0.4%	1.5%	1.6%	0.9%
4	11.2%	2.6%	1.8%	1.5%	2.2%	9.6%	(2.9%)	(1.9%)
5	(3.9%)	(1.3%)	0.9%	0.9%	0.8%	(3.5%)	2.9%	0.2%

Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES.

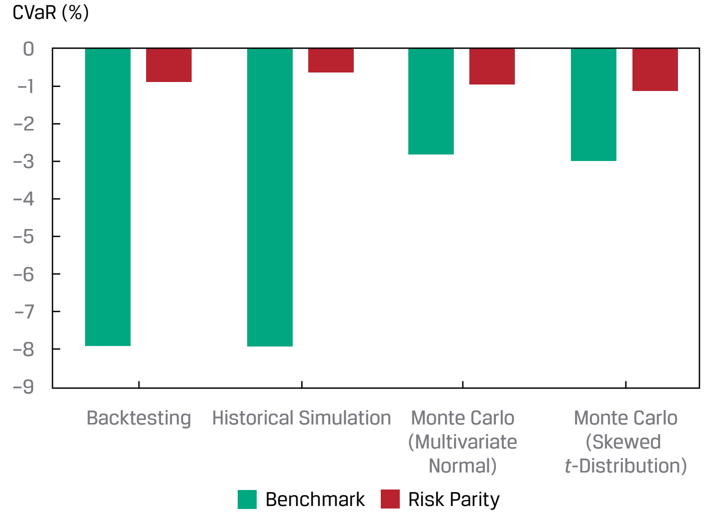
Turning to the performance and risk profiles of our investment strategies, shown in Panel A of Exhibit 24, we note that the Sharpe ratio appears insensitive to any of the particular simulation methods used by consistently suggesting that the risk parity allocation strategy outperforms the benchmark strategy. Downside risk (expressed as CVaR), however, appears quite sensitive to the choice of simulation approach for the BM strategy, but not very sensitive for the RP strategy (Panel B). If we focus on the BM strategy, the CVaR results from historical simulation and rolling-window backtesting resemble each other very closely. The CVaR results of both (multivariate skewed *t*- and multivariate normal) Monte Carlo simulations are also very similar: Both underestimate the downside risk of the BM strategy. This finding suggests that additional sensitivity analyses should be run with different functional forms for the factor return distributions.

Exhibit 24: Comparing Simulation Methods with Backtesting

A. Sharpe Ratio



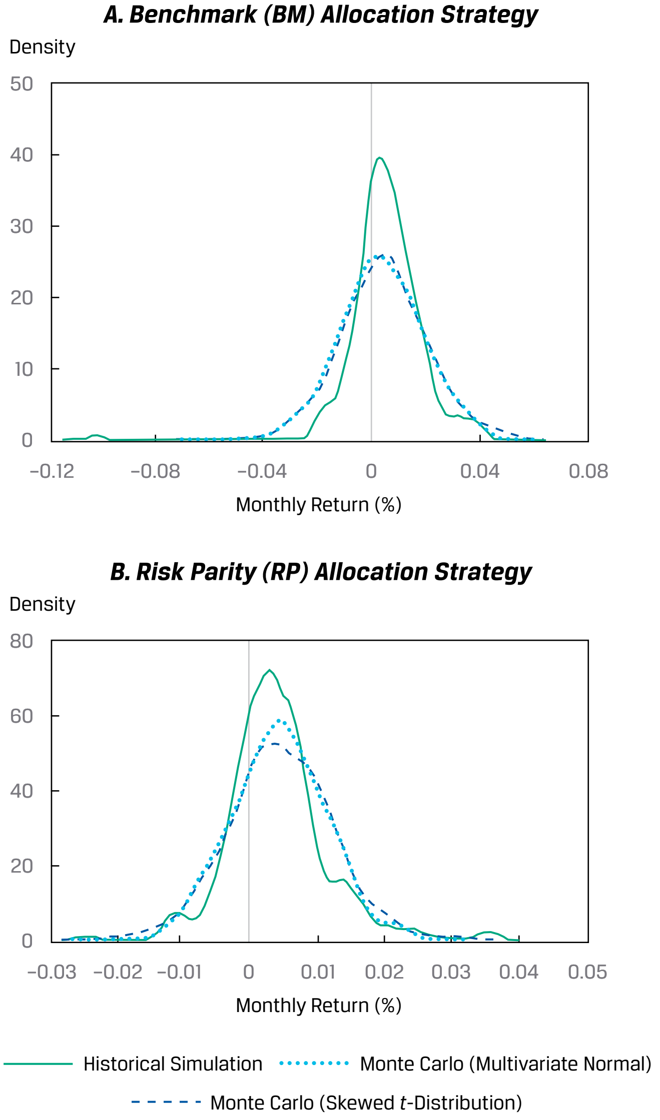
B. Conditional Value-at-Risk



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES.

Estimated probability density plots, in Panel A of Exhibit 25, show that the difference between the historical simulation and the two Monte Carlo methods is rather large for the BM strategy. Given the negative skewness and excess kurtosis of the BM strategy's returns, which is apparent from the shape of the historical simulation return distribution, it is not surprising that the two Monte Carlo simulations fail to account sufficiently for this left-tail risk property. Conversely, because the distribution of the RP strategy's returns is relatively symmetric and without much excess kurtosis, all three simulation methods provide a fairly similar picture (Panel B).

Exhibit 25: Estimated Distribution Plots for BM and RP Strategies Using Three Different Simulations



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES.

EXAMPLE 8

Simulating the Performance of Factor Allocation Strategies

Earlier, Sarah Koh presented her team's backtesting results for the factor-based allocation strategies being considered by an important client, SWF Fund. Now, while presenting the simulation results for these same strategies, SWF Fund's investment committee asks Koh the following questions:

1. The following are caveats regarding the use of rolling-window backtesting in assessing investment strategies *except*:

   A. this technique implicitly assumes that the same pattern of past performance is likely to repeat itself over time.

   B. this technique may not fully account for the dynamic nature of financial markets and potentially extreme downside risks.

   C. this technique is intuitive, because it mimics how investing is done in reality—that is, forming ideas, testing strategies, and implementing periodically.

Solution:

C is correct, because it is not a caveat in using rolling-window backtesting. A and B are incorrect because they are caveats in the use of this technique.

2. Which of the following situations is *most likely* to involve data snooping?

   A. A researcher performs rolling-window backtesting of a new momentum strategy using 20 years of point-in-time (PIT) data from the United States. She cross-validates results by similarly analyzing PIT data from the following markets: mainland China, Asia ex-Japan, Europe, the United Kingdom, and Canada.

   B. A researcher tries many different modeling techniques, backtesting each of them, and then picking the best-performing model without accounting for model selection bias.

   C. A researcher sets a relatively high hurdle, a *t*-statistic greater than 3.0, for assessing whether a newly discovered factor is statistically significant.

Solution:

B is correct, because this situation most likely involves data snooping. A and C are incorrect because these are approaches to avoiding data snooping.

3. Which of the following situations is *least likely* to involve scenario analysis?

   A. Simulating the performance and risk of investment strategies by first using stocks in the Nikkei 225 Index and then using stocks in the TOPIX 1000 Index.

   B. Simulating the performance and risk of investment strategies in both "trade agreement" and "no-trade-agreement" environments.

   C. Simulating the performance and risk of investment strategies in both high-volatility and low-volatility environments.

Solution:

A is correct, because there is no structural break or different structural regime. B and C are incorrect because they involve structural breaks/different structural regimes and thus represent different scenarios.

4. Which one of the following statements concerning historical simulation and Monte Carlo simulation is *false*?

   A. Historical simulation randomly samples (with replacement) from the past record of asset returns, where each set of past monthly returns is equally likely to be selected.

B. Neither historical simulation nor Monte Carlo simulation makes use of a random number generator.

C. Monte Carlo simulation randomly samples from an assumed multivariate joint probability distribution in which the past record of asset returns is used to calibrate the parameters of the multivariate distribution.

Solution:

B is correct, because this statement is false. A and C are incorrect because they are true statements about historical and Monte Carlo simulation, respectively.

5. Which one of the following statements concerning Monte Carlo simulation is *false*?

A. When simulating multiple assets (factors) whose returns are correlated, it is crucial to specify a multivariate distribution rather than modeling each asset on a standalone basis.

B. Regression and distribution-fitting techniques are used to estimate the parameters underlying the statistical distributions of the key decision variables.

C. The Monte Carlo simulation process is deterministic and non-random in nature.

Solution:

C is correct, because this statement is false. A and B are incorrect because they are true statements about Monte Carlo simulation.

6. Which of the following situations concerning simulation of a multifactor asset allocation strategy is *most likely* to involve sensitivity analysis?

A. Changing the specified multivariate distribution assumption from a normal to a skewed *t*-distribution to better account for skewness and fat tails

B. Splitting the rolling window between periods of recession and non-recession

C. Splitting the rolling window between periods of high volatility and low volatility

Solution:

A is correct, because this choice represents sensitivity analysis. B and C are incorrect because these choices represent scenario analysis.

SUMMARY

In this reading, we have discussed how to perform rolling-window backtesting—a widely used technique in the investment industry. We also described how to use scenario analysis and simulation along with sensitivity analysis to supplement backtesting, so investors can better account for the randomness in data that may not be fully captured by backtesting.

- The main objective of backtesting is to understand the risk–return trade-off of an investment strategy by approximating the real-life investment process.
- The basic steps in rolling-window backtesting are specifying the investment hypothesis and goals, determining the rules and processes behind an investment strategy, forming an investment portfolio according to those rules, rebalancing the portfolio periodically, and computing the performance and risk profiles of the strategy.
- In the rolling-window backtesting methodology, researchers use a rolling-window (or walk-forward) framework, fit/calibrate factors or trade signals based on the rolling window, rebalance the portfolio periodically, and then track the performance over time. Thus, rolling-window backtesting is a proxy for actual investing.
- Analysts need to pay attention to several behavioral issues in backtesting, including survivorship bias and look-ahead bias.
- Asset (and factor) returns are often negatively skewed and exhibit excess kurtosis (fat tails) and tail dependence compared with a normal distribution. As a result, standard rolling-window backtesting may be unable to fully account for the randomness in asset returns, particularly on downside risk.
- Financial data often face structural breaks. Scenario analysis can help investors understand the performance of an investment strategy in different structural regimes.
- Historical simulation is relatively straightforward to perform but shares pros and cons similar to those of rolling-window backtesting. For example, a key assumption these methods share is that the distribution pattern from the historical data is sufficient to represent the uncertainty in the future. Bootstrapping (or random draws with replacement) is often used in historical simulation.
- Monte Carlo simulation is a more sophisticated technique than historical simulation. In Monte Carlo simulation, the most important decision is the choice of functional form of the statistical distribution of decision variables/return drivers. Multivariate normal distribution is often used in investment research, owing to its simplicity. However, a multivariate normal distribution cannot account for negative skewness and fat tails observed in factor and asset returns.
- Sensitivity analysis, a technique for exploring how a target variable and risk profiles are affected by changes in input variables, can further help investors understand the limitations of conventional Monte Carlo simulation (which typically assumes a multivariate normal distribution as a starting point). A multivariate skewed *t*-distribution considers skewness and kurtosis but requires estimation of more parameters and thus is more likely to suffer from larger estimation errors.

PRACTICE PROBLEMS

The following information relates to questions 1-8

Kata Rom is an equity analyst working for Gimingham Wealth Partners (GWP), a large investment advisory company. Rom meets with Goran Galic, a Canadian private wealth client, to explain investment strategies used by GWP to generate portfolio alpha for its clients.

Rom states that GWP is recognized in the Canadian investment industry as a leading factor-based value portfolio manager and describes how GWP creates relevant investment strategies and explains GWP's backtesting process. Rom notes the following:

Statement 1 Using historical data, backtesting approximates a real-life investment process to illustrate the risk–return tradeoff of a particular proposed investment strategy.

Statement 2 Backtesting is used almost exclusively by quantitative investment managers and rarely by fundamental investment managers, who are more concerned with information such as forward estimates of company earnings, macroeconomic factors, and intrinsic values.

Galic, who is 62 years old, decides to allocate C$2 million (representing 10% of his net worth) to an account with GWP and stipulates that portfolio assets be restricted exclusively to domestic securities. Although GWP has not backtested its strategies with such a restriction, it has backtested its strategies using a global index that includes domestic securities. Rom shows the following risk measures to Galic for three factor portfolios.

Exhibit 1: Downside Risk Measures for Model Factors

Risk Measure	Factor 1	Factor 2	Factor 3
Value at risk (VaR) (95%)	(6.49%)	(0.77%)	(2.40%)
Conditional VaR (CVaR) (95%)	(15.73%)	(4.21%)	(3.24%)
Maximum drawdown	35.10%	38.83%	45.98%

Galic asks Rom, "What happens if the future is different from the past?" Rom gives the following replies:

Statement 3 Although backtesting can offer some comfort, you are correct that it does have a weakness: Backtesting generally does not capture the dynamic nature of financial markets and in particular may not capture extreme downside risk.

Statement 4 As a result, we have captured extreme downside risk and the dynamic nature of financial markets by using the Value-at-Risk and Conditional Value-at-Risk measures.

In an effort to make Galic fully aware of the risks inherent in GWP's strategies, Rom describes a recent study that investigated the return distributions of value and momentum factors that GWP uses to construct portfolios. The study found that these distributions were non-normal based on their negative skewness, excess kurtosis, and tail dependence. Rom indicated that investment strategies based on this type of data are prone to significantly higher downside risk. Rom informs Galic that GWP also uses a technique commonly referred to as scenario analysis to examine how strategies perform in different structural regimes. Exhibit 2 compares the performance of two of GWP's factor allocation strategies in different regimes:

Exhibit 2: Scenario Analysis Using the Sharpe Ratio

Strategy/Regime	High Volatility	Low Volatility	Recession	Non-recession
Strategy I	0.88	0.64	0.20	1.00
Strategy II	1.56	1.60	1.76	1.52

Galic is surprised to see that some of the backtest results are unfavorable. He asks, "Why has GWP not considered strategies that perform better in backtesting?" Galic recently met with Fastlane Wealth Managers, who showed much better performance results. The portfolio manager at Fastlane told Galic that the company selects the top-performing strategies after performing thousands of backtests.

1. Which of Rom's statements concerning backtesting is correct?

   A. Only Statement 1

   B. Only Statement 2

   C. Both Statement 1 and Statement 2

2. Which key parameter needs to be changed for a new backtest that includes Galic's restrictions?

   A. Start and end dates

   B. Consideration of transaction costs

   C. Investment universe

3. Galic's concern embedded in the question "What happens if the future is different from the past?" is a problem most relevant for which investment strategy evaluation technique?

   A. Sensitivity analysis

   B. Backtesting

   C. Monte Carlo simulation

4. Which of the following conclusions of Exhibit 1 is *least* likely to be true?
   A. 5% of the time, losses from Factor 1 would be at least 6.49%.
   B. When the VaR is exceeded in Factor 1, we should expect an average loss of 15.73%.
   C. 5% of the time, losses from Factor 2 are likely to be worse than losses from Factor 1.

5. Based on the statistical study performed by GWP, which of the following represents a suggested course of action if GWP were to conduct Monte Carlo simulation analyses on the factor strategies?
   A. Inverse transformation
   B. Bootstrapping
   C. Sensitivity analysis

6. Based on Exhibit 1, which factor has the smallest downside risk as measured by the weighted average of all losses that exceed a threshold?
   A. Factor 1
   B. Factor 2
   C. Factor 3

7. The approach used by Fastlane Wealth Managers *most likely* incorporates:
   A. risk parity.
   B. data snooping.
   C. cross-validation.

8. Comparing the two strategies in Exhibit 2, the *best* risk-adjusted performance is demonstrated by:
   A. Strategy II in periods of low volatility and recession.
   B. Strategy I in periods of high volatility and non-recession.
   C. Strategy II in periods of high volatility and non-recession.

The following information relates to questions 9-16

Emily Yuen is a senior analyst for a consulting firm that specializes in assessing equity strategies using backtesting and simulation techniques. She is working with an assistant, Cameron Ruckey, to develop multifactor portfolio strategies based on nine factors common to the growth style of investing. To do so, Yuen and Ruckey plan to construct nine separate factor portfolios and then use them to create factor-weighted allocation portfolios.

Yuen tasks Ruckey with specifying the investment universe and determining the availability of appropriate reporting data in vendor databases. Ruckey selects a vendor database that does not provide point-in-time data, so he adjusts the database to include point-in-time constituent stocks and a reporting lag of four months.

Next, Yuen and Ruckey run initial backtests on the nine factor portfolios, calculating performance statistics and key metrics for each. For backtesting purposes, the portfolios are rebalanced monthly over a 30-year time horizon using a rolling-window procedure.

Yuen and Ruckey consider a variety of metrics to assess the results of the factor portfolio backtests. Yuen asks Ruckey what can be concluded from the data for three of the factor strategies in Exhibit 1:

Exhibit 1: Backtest Metrics for Factor Strategies

	Factor 1	Factor 2	Factor 3
VaR (95%)	(3.9%)	(1.3%)	(8.4%)
Standard deviation of returns	2.1%	1.2%	4.6%
Maximum drawdown	27.2%	8.3%	59.7%

Ruckey tells Yuen the following:

Statement 1 We do not need to consider maximum drawdown, because standard deviation sufficiently characterizes risk.

Statement 2 Factor 2 has the highest downside risk.

From her professional experience Yuen knows that benchmark and risk parity factor portfolios, in which factors are equally weighted and equally risk weighted, respectively, are popular with institutional and high-net-worth clients. To gain a more complete picture of these investment strategies' performance, Yuen and Ruckey design a Benchmark Portfolio (A) and a Risk Parity Portfolio (B), and then run two simulation methods to generate investment performance data based on the underlying factor portfolios, assuming 1,000 simulation trials for each approach:

Approach 1 Historical simulation

Approach 2 Monte Carlo simulation

Yuen and Ruckey discuss the differences between the two approaches and then design the simulations, making key decisions at various steps. During the process, Yuen expresses a number of concerns:

Concern 1: Returns from six of the nine factors are correlated.

Concern 2: The distribution of Factor 1 returns exhibits excess kurtosis and negative skewness.

Concern 3: The number of simulations needed for Approach 1 is larger than the size of the historical dataset.

For each approach, Yuen and Ruckey run 1,000 trials to obtain 1,000 returns for Portfolios A and B. To help understand the effect of the skewness and excess kurtosis observed in the Factor 1 returns on the performance of Portfolios A and B, Ruckey suggests simulating an additional 1,000 factor returns using a multivariate skewed Student's *t*-distribution, then repeating the Approach 2 simulation.

9. Following Ruckey's adjustments to the initial vendor database, backtested returns will *most likely* be subject to:

   A. stale data.

   B. data snooping

   C. p-hacking

10. Based on Exhibit 1, Ruckey should conclude that:

    A. Factor Strategy 3 has the highest portfolio turnover.

    B. Factor Strategy 2 has less downside risk than Strategy 3.

    C. Factor Strategy 2 has the highest returns.

11. Which of Ruckey's statements about Exhibit 1 is <u>in</u>correct?

    A. Only Statement 1

    B. Only Statement 2

    C. Both Statement 1 and Statement 2

12. Simulation Approach 1 (historical simulation) differs from Approach 2 (Monte Carlo simulation) in that:

    A. it is deterministic.

    B. a functional form of the statistical distribution for each decision variable needs to be specified.

    C. it assumes that sampling the returns from the actual data provides sufficient guidance about future asset returns.

13. To address Concern 1 when designing Approach 2, Yuen should:

    A. model each factor or asset on a standalone basis.

    B. calculate the 15 covariance matrix elements needed to calibrate the model.

    C. specify a multivariate distribution rather than modeling each factor or asset on a standalone basis.

14. Based on Concern 2, the Factor 1 strategy is *most likely* to:

    A. be favored by risk-averse investors.

    B. generate surprises in the form of negative returns.

    C. have return data that line up tightly around a trend line.

15. To address Concern 3 when designing Approach 1, Yuen should:

    A. add monthly return observations to the dataset using interpolation.

    B. randomly sample from the historical returns with replacement.

    C. choose the multivariate normal distribution as the initial functional form.

16. The process Ruckey suggests to better understand how the performance of Portfolios A and B using Approach 2 is affected by the distribution of Factor 1 returns is *best* described as:

    A. data snooping.

    B. sensitivity analysis.

    C. inverse transformation.

SOLUTIONS

1. A is correct. Statement 1 is correct because the main objective of backtesting is to understand the risk–return tradeoff of an investment strategy by approximating the real-life investment process.

   B is incorrect because Statement 2 is inaccurate. Although backtesting fits quantitative and systematic investment styles more naturally, it has also been heavily used by fundamental managers.

   C is incorrect because Statement 2 is not accurate. Backtesting, used in quantitative and systematic investment styles, is also heavily used by fundamental managers.

2. C is correct. Investment universe represents the securities in which a strategy can potentially invest. Galic's restriction to exclusively own domestic securities means the investment universe of a backtest for a strategy for Galic's account should use a domestic rather than global investment universe.

   A is incorrect. Galic's restriction to domestic securities does not affect the start and end dates for a backtest.

   B is incorrect. Galic's restriction to domestic securities does not change the inclusion of transaction costs in the study.

3. B is correct. An implicit assumption of backtesting is that past returns are a guide to future asset returns.

   A is incorrect. Sensitivity analysis refers to modifying assumptions such as probability distributions of key variables in a Monte Carlo simulation, which is a non-deterministic evaluation technique that does not use historical data.

   C is incorrect. Monte Carlo simulation is a non-deterministic evaluation technique that does not use historical data.

4. C is correct. The VaR metrics in Exhibit 1 show that 5% of the time, losses will be at least 6.49% and 0.77%, respectively, for Factor 1 and Factor 2. The CVaR metrics in Exhibit 1 show that the weighted average of all loss outcomes that exceed the VaR loss are 15.73% and 4.21% for Factor 1 and Factor 2, respectively. Thus, A is true because it correctly defines VaR, and B is true because it correctly defines CVaR, whereas C is untrue because both VaR and CVaR are lower for Factor 2 than Factor 1.

5. C is correct. Performing sensitivity analysis represents best practice given these characteristics, because the user could test different probability distributions that relax the assumptions of the normal distribution, for example.

   A is incorrect. Inverse transformation is a method of random observation generation, often used in simulation.

   B is incorrect. Bootstrapping refers to random sampling with replacement, often used in historical simulation.

6. C is correct. Exhibit 1 presents three downside risk measures: VaR, CVaR, and maximum drawdown. Conditional VaR is defined as the weighted average of all loss outcomes in the return distribution that exceed the VaR loss. Thus, CVaR is a more comprehensive measure of tail loss than VaR. Based on Exhibit 1, the factor with the smallest downside risk based on CVaR is Factor 3.

7. B is correct. The fact that the two firms' investment performance results differ over similar time horizons using the same data and factors may be the result of selection bias. Data snooping is a type of selection bias. Fastlane Wealth Manag-

ers is most likely selecting the best-performing modeling approach and publishing its results (i.e., data snooping).

A is incorrect because risk parity is a portfolio construction technique that accounts for the volatility of each factor and the correlations of returns among all factors to be combined in the portfolio. It is not regarded as selection bias.

C is incorrect because cross-validation is a technique used in the machine learning field, as well as in backtesting investment strategies, to partition data for model training and testing. It is not considered selection bias.

8. A is correct. Using the Sharpe ratio, the best risk-adjusted relative performance can be determined by comparing the sensitivity of the two strategies under differing macroeconomic regimes: recession versus non-recession and high volatility versus low volatility. The best risk-adjusted return will exhibit the highest Sharpe ratio. Strategy II demonstrates higher risk-adjusted returns compared with Strategy I under all four macroeconomic conditions, particularly in periods of low volatility, when the Sharpe ratio outperformance is 0.96, and recessions, when the Sharpe ratio outperformance is 1.56.

9. A is correct. A reporting lag of four months is likely to introduce stale data into the backtest because many large-capitalization companies report earnings within 30–50 days of quarter end. Although assuming four months (120 days) of reporting lag will eliminate a source of look-ahead bias, it introduces a new problem (i.e., stale data).

   B and C are incorrect. Data snooping and p-hacking refer to the same problem: a flawed approach to using data to make decisions. Data snooping and p-hacking are not characteristics of data, nor can they be added to a dataset by making an adjustment.

10. B is correct. Both VaR and maximum drawdown are downside risk measures, and both measures are lower for Strategy 2 than Strategy 3.

    A is incorrect. We cannot deduce portfolio turnover from the metrics provided in Exhibit 1.

    C is incorrect. We cannot deduce returns from the metrics provided in Exhibit 1.

11. C is correct. Both statements are incorrect. Statement 1 is incorrect because maximum drawdown and standard deviation are different measures. Maximum drawdown is typically used to represent downside risk, because it is the minimum cumulative return observed. Standard deviation is a measure of volatility. Although the two measures may be correlated, they are not substitutes for each other. Statement 2 is incorrect because two downside risk measures are presented: VaR and maximum drawdown. Factor Strategy 2 has the lowest reading for both measures, indicating that it has the *least* downside risk among the three strategies presented in Exhibit 1.

12. C is correct. Approach 1 is a historical simulation and assumes that past asset returns provide sufficient guidance about future asset returns.

    A is incorrect because both approaches are non-deterministic and random in nature. Approach 1 is a historical simulation, and Approach 2 is a Monte Carlo simulation.

    B is incorrect because Approach 1 is a historical simulation and each random variable of interest (key driver and/or decision variable) is randomly drawn from historical data. A functional form of the statistical distribution of returns for each decision variable needs to be specified for a Monte Carlo simulation, which is Approach 2.

13. C is correct. Approach 2 is a Monte Carlo simulation. The returns of Portfolios A

and B are driven by the returns of the nine underlying factor portfolios (based on nine common growth factors). In the case of asset or factor allocation strategies, the returns from six of the nine factors are correlated, and therefore it is necessary to specify a multivariate distribution rather than modeling each factor or asset on a standalone basis.

A is incorrect because Approach 2 is a Monte Carlo simulation to generate investment performance data for the nine underlying factor portfolios. The returns of six of the nine factors are correlated, which means specifying a multivariate distribution rather than modeling each factor or asset on a standalone basis.

B is incorrect because the analyst should calculate the elements of the covariance matrix for all factors, not only the correlated factors. Doing so entails calculating 36, not 15, elements of the covariance matrix. Approach 2 is a Monte Carlo simulation using the factor allocation strategies for Portfolios A and B for the nine factor portfolios, the returns of which are correlated, which means specifying a multivariate distribution. To calibrate the model, a few key parameters need to be calculated: the mean, the standard deviation, and the covariance matrix. For 9 assets, we need to estimate 9 mean returns, 9 standard deviations, and $\frac{9 \times (9-1)}{2} = 36$ elements of the covariance matrix. Assuming just the 6 correlated assets, the calculation is $\frac{6 \times (6-1)}{2} = 15$.

14. B is correct. The distribution of Factor 1 returns exhibits excess kurtosis and negative skewness (relative to the normal distribution). The excess kurtosis implies that these strategies are more likely to generate surprises, meaning extreme returns, whereas the negative skewness suggests those surprises are more likely to be negative (than positive).

    A is incorrect because risk-averse investors are more likely to prefer distribution properties such as positive skew (higher probability of positive returns) and lower to moderate kurtosis (lower probability of extreme negative surprises). The distribution of Factor 1 returns exhibits excess kurtosis and negative skewness.

    C is incorrect because the distribution of Factor 1 returns exhibits excess kurtosis and negative skewness. The joint distribution of such returns is rarely multivariate normal—so, typically the means and variances of these returns and the correlations between them are insufficient to describe the joint return distribution. In other words, the return data do not line up tightly around a trend line because of fat tails and outliers.

15. B is correct. Random sampling with replacement, also known as bootstrapping, is often used in historical simulations because the number of simulations needed is often larger than the size of the historical dataset. Because Approach 1 is a historical simulation and Concern 3 notes that the number of simulations needed is larger than the size of the historical dataset, bootstrapping should be used.

    A is incorrect because this approach would result in creating observations that do not exist in the historical record. Doing so would violate the assumption and procedures of historical simulation.

    C is incorrect because choosing the multivariate normal distribution as the initial functional form is typically done in a Monte Carlo simulation (Approach 2), not in a historical simulation (Approach 1). Historical simulation randomly samples from the historical dataset by drawing a number from a uniform distribution so that there is equal probability of being selected. Choice of distribution would not address the concern about the size of the dataset.

16. B is correct. Sensitivity analysis can be implemented to help managers understand how the target variable (portfolio returns) and risk profiles are affected by changes in input variables. Approach 2 is a Monte Carlo simulation, and the results depend on whether the multivariate normal distribution is the correct

functional form or a reasonable proxy for the true distribution. Because this information is almost never known, sensitivity analysis using a multivariate skewed Student's *t*-distribution helps to account for empirical properties such as the skewness and the excess kurtosis observed in the underlying factor return data.

A is incorrect. Data snooping is the subconscious or conscious manipulation of data in a way that produces a statistically significant result (i.e., a *p*-value that is sufficiently small or a *t*-statistic that is sufficiently large to indicate statistically significance).

C is incorrect. The inverse transformation method is the process of converting a randomly generated number into a simulated value of a random variable.