Federal Reserve Bank
of Chicago

Third Quarter 1999

Economic perspectives

Economic perspectives

Economic Perspectives is published by the Research Department of the Federal Reserve Bank of Chicago. The views expressed are the authors' and do not necessarily reflect the views of the Federal Reserve Bank of Chicago or the Federal Reserve System.

Single-copy subscriptions are available free of charge. Please send requests for single- and multiple-copy subscriptions, back issues, and address changes to the Public Information Center, Federal Reserve Bank of Chicago, P.O. Box 834, Chicago, Illinois 60690-0834, telephone 312-322-5111 or fax 312-322-5515.

Economic Perspectives and other Bank publications are available on the World Wide Web at http://www.frbchi.org.

ISSN 0164-0682

Contents

Third Quarter 1999, Volume XXIII, Issue 3

State budgets and the business cycle: Implications for the federal balanced budget amendment debate

Leslie McGranahan

Introduction and summary

A proposal to amend the U.S. Constitution to require that the federal budget be balanced has been a part of the national debate for over 25 years. Following its inclusion as one of the central planks of the Republican Contract with America in 1994, the balanced budget amendment became a prominent item on the congressional agenda. The amendment easily passed the House by a vote of 300 to 132 in January 1995, but failed to achieve the two-thirds majority required in the Senate to send it back to the states. Since the proposal's most recent failure in the Senate, by one vote on March 4, 1997, it has been a less important agenda item because of the strength of the economy and the surplus in the federal budget. However, the issue is by no means dead. In January 1999, the amendment was again proposed in the House with the cosponsorship of 117 representatives.

Balanced budget amendment supporters frequently cite the experience of the states, most of which have statutory or constitutional balanced budget restrictions.[1] In this article, I question how the state experience with balanced budget restrictions can inform the federal debate on a balanced budget amendment. First, I address how the longstanding state restrictions compare with those contemplated at the federal level. I then investigate how state government revenues, expenditures, debt issuance, and asset holdings have responded to changes in the states' economic conditions, as measured by the unemployment rate, during the last two decades. I use regression analysis to ask how, controlling for a time trend and state fixed effects, state finances have reacted to fiscal year state unemployment rates from 1977 to 1997. I further question whether similar responses on the part of the federal government would be either feasible or prudent.

In my investigation of how state finances respond to business cycle conditions, I discover that states use four main mechanisms to maintain budget balances during downturns: they issue more short- and long-term debt; they rely more heavily on the federal government for funds while giving less to local governments; they increase tax rates; and they lower capital spending. This is not a feasible policy combination for the federal government for a number of reasons. Most importantly, the provisions of the balanced budget amendment would not allow the federal government to issue any new debt without a legislative supermajority. In this way, the federal balanced budget amendment differs significantly from the restrictions in place in the states. While the states use the issuance of debt as an important safety valve, this option would not be available to the federal government.

Of course, the opportunity to receive more from a higher level of government would also not be available to the federal government. However, the federal government could follow the states' lead by transferring less money to the states during difficult times. This would reverse the current relationship between federal government intergovernmental spending and the business cycle and would make it more difficult for the state governments to balance their budgets. Importantly, this suggests that one of the reasons that the states are able to balance their budgets is that the federal government does not.

The federal government could follow the states by increasing tax rates during economic downturns. This would be an unpopular policy for two main reasons. First, tax increases are always unpopular and difficult to pass. Second, unlike the state governments, the federal government is responsible for the condition

Leslie McGranahan is an economist in the Economic Research Department of the Federal Reserve Bank of Chicago. *The author would like to thank Loula Sassaris for research assistance and colleagues at the Federal Reserve Bank of Chicago for help and comments.*

of the macroeconomy. Tax increases during recessions would further depress disposable incomes and consumption and could prolong downturns.

The other state behavior open to the federal government would be to decrease capital spending during economic downturns. States get a lot of leverage out of their ability to cut capital spending during difficult times; my results show that this is among the most pronounced state responses to a deteriorating economic situation. The federal government may be unwilling to follow the states' lead by cutting capital spending during recessions because the bulk of federal capital spending, over 80 percent, is in the area of national defense (U.S. Government, Office of Management and Budget, 1999). By contrast, the majority of state capital spending is on highways (57 percent) and institutions of higher education (14 percent) (U.S. Department of Commerce, Bureau of the Census, 1977–87 and 1988–97). Whether it is prudent for the federal government to structure defense capital spending to maintain budget balance during downturns is an open question.

Because of the differences in the proposed federal balanced budget amendment and the measures in place in the states and the different responsibilities of the federal versus state governments, none of the four methods used by state governments during economic downturns is an obvious choice for the federal government. In summary, my results suggest that the ability of the states to function under their current balanced budget restrictions should not be used to argue in favor of the balanced budget amendment most recently proposed in Congress. However, this does not necessarily imply that other reasons advanced in favor of a balanced budget amendment are invalid or that the amendment should not be justified on other grounds.

Comparing state and federal balanced budget requirements

The provisions of the proposed federal balanced budget amendment are quite basic. The amendment as voted on in 1997 simply states that "[t]otal outlays for any fiscal year shall not exceed total receipts for that fiscal year, unless three-fifths of the whole number of each House of Congress shall provide by law for a specific excess of outlays over receipts by a roll-call vote." Additional provisions require a three-fifths majority to increase the debt limit or to increase revenues (U.S. Senate, 1997).

The amendment does not provide for separate funds to finance capital projects and, therefore, in the absence of a super-majority vote, does not allow the government to issue any new debt. In addition, the amendment does not provide for a reserve fund that can be used to carry over surpluses from one year to the next. Instead, surpluses that were neither spent nor returned to citizens would be used to reduce the existing debt. This arises from the provisions that outlays must be financed by total receipts from the same fiscal year—it does not allow for the use of receipts from previous years. Both of these features would be in contrast to the provisions in the states. In short, the amendment simply requires that the budget be balanced every year.

State balanced budget restrictions are far more complex than the federal proposal. There is no prototypical requirement at state level; each state has a unique set of provisions. However, the following state provisions are comparable to the federal proposal: balanced budget requirements, restrictions on deficit carryovers, and restrictions on long-term debt issuance.

Before addressing how these three types of restrictions interact to affect state behavior, it is important to briefly explain the role of capital budgeting in the states. Most states have capital budgets that are separate from their operating budgets.[2] The construction of new facilities and the repair, maintenance, remodeling, and rehabilitation of existing facilities are funded separately.[3] One important feature that distinguishes state balanced budget requirements from those at the federal level is that most of these requirements only mandate that the operating budget be balanced. In cases where the capital budget also needs to be balanced, proceeds from the issuance of debt are counted as revenues. Therefore, the balanced budget restrictions do not stop states from issuing debt. This contrasts with the federal proposal, which explicitly excludes receipts derived from borrowing from government revenues. The ability of states to borrow for capital projects reconciles the common perception that states have balanced budgets with a thriving and substantial municipal bond market.

Submitting, passing, or signing a balanced budget

When commentators write that most states have balanced budget restrictions, they are usually referring to constitutional or statutory provisions that require that the governor must submit, the legislature must pass, or the governor must sign a balanced budget. These provisions do not explicitly require that the year-end budget end up balanced, but rather that the budget as proposed, passed, or signed be balanced in expectation. For example, the Illinois constitution requires that the governor submit and the legislature pass a balanced budget. The document states, "[t]he Governor shall prepare and submit to the General

Assembly, at a time prescribed by law, a State budget for the ensuing fiscal year. … Proposed expenditures shall not exceed funds *estimated* to be available for the fiscal year as shown in the budget." It further states that "[a]ppropriations for a fiscal year shall not exceed funds *estimated* by the General Assembly to be available during that year" (italics added) (State of Illinois, 1970, Article 8, Section 2). Note that in both cases expenditures cannot exceed estimated revenues.

Deficit carryover provisions

In the event that circumstances change during the year and a budget that was expected or estimated to be balanced was not, state provisions either allow or do not allow deficits to be carried over from one fiscal year to the next. If the state does not allow deficits to be carried over, the state must either cut spending or increase revenues to eradicate the deficit by fiscal year-end. Such deficit carryover provisions represent the teeth of the balanced budget requirements, because they prohibit the state from issuing debt to finance a shortfall. The National Conference of State Legislatures reports that 13 states have no restriction on carrying over a deficit and a total of 21 may carry over a deficit if necessary (Snell, 1999). Illinois is one of the states allowed to carry over deficits. The Illinois constitution states that "[s]tate debt may be incurred by law in an amount not exceeding 15 percent of the State's appropriations for that fiscal year to meet deficits caused by emergencies or failures of revenue" (State of Illinois, 1970, Article 9, Section 9).[4] Note that all states do allow surpluses to be carried over from one year to the next and 45 states have special "rainy day funds" for surplus carryovers (Eckl, 1998).

State long-term debt provisions

The final parts of states' budget restrictions are provisions limiting their ability to issue long-term debt. Nearly all long-term debt is used to finance specific capital projects in conjunction with the state's capital budget. While federal Treasury bonds, notes, and bills are very general in nature, most state government debt is very specific and is issued to benefit particular capital projects. State debt can be backed by either the full faith and credit or the taxing power of the government, and can be redeemed from general revenues or be nonguaranteed and be backed by specific income streams.

Most states have a restriction limiting the issuance of long-term debt. Some state constitutions require that debt cannot be issued until it receives majority support in a statewide referendum; in some states debt can only be issued up to a prespecified limit; and other states allow no debt to be issued at all.[5] However, state courts have interpreted these constitutional requirements as only applying to debt backed by the full faith and credit of the government. As a result, states can issue nonguaranteed debt limited only by the constraints of the capital market. In fact, despite restrictions on long-term debt that in some cases seem quite severe, in every year since 1977 every state has issued some long-term debt.

In sum, the restrictions on the states are far more lenient than that contemplated for the federal government. In particular, all states can and do issue long-term debt and many states can issue debt to finance deficits.

Nonetheless, the states' experience with budget restrictions is frequently used to support balanced budget restrictions at the federal level. For example, Michigan's Governor John Engler in his State of the State Address in 1997 said, "I support the balanced budget amendment and so do Michigan voters. When Congress takes up this historical amendment next month, I urge them to pass it and submit it to the states. I invite this legislature to join the debate, call upon your colleagues in Congress to act and help the federal budget look more like Michigan's budget—balanced" (Engler, 1997). Similarly, in his 1997 State of the State address Oklahoma Governor Frank Keating stated that "We Oklahomans know the wisdom of a constitutional mandate for fiscal common sense. Let's send some Oklahoma values to Washington by being the first to ratify this vital amendment" (Keating, 1997).

While the current state restrictions and the contemplated federal restrictions are quite different, the general perception that states are more fiscally responsible is warranted. States do a better job on two dimensions. First, they have a lower level of overall debt relative to their financial obligations. Between 1977 and 1997, net interest payments on the federal debt averaged 12.7 percent of outlays and 15.0 percent of receipts (U.S. Government, Office of Management and Budget, 1999), while state interest payments averaged 3.7 percent of expenditures and 3.4 percent of revenues (U.S. Department of Commerce, Bureau of the Census, 1977–87 and 1988–97).[6] Similarly, gross federal debt outstanding averaged 2.3 times outlays and 2.7 times receipts, while state gross debt outstanding averaged 0.5 times revenues and 0.6 times outlays over the same period. Second, the states do a better job of smoothing over the business cycle. A 1 percentage point increase in the state unemployment rate increases the average state's budget deficit (expenditures – revenues) by $23 per capita or about 9 percent (relative to the mean), while a 1 percentage

point increase in the national unemployment rate increases the federal government deficit by $134 or about 16 percent.

Next, I investigate how state budget items respond to business cycle conditions. If a federal balanced budget amendment were to pass, the federal government would need to find ways to either raise additional funds or cut expenditures to compensate for the decline in tax revenues that accompanies downturns. The assumption that the federal government could mimic the cyclical behavior of the states is implicit in the argument that state experience is a valid example for the federal government. I ask what the states do and whether the state experience could or should be mimicked by the federal government.

Data and methodology

To look at how state finances change over the business cycle, I need data on both business cycle conditions within a state and on various attributes of state government finances.

Measuring the business cycle

To measure business conditions in the state, I use the average monthly state unemployment rate during the fiscal year for which the state finance data are collected. For the most part, the analysis focuses on state fiscal years (FY) 1977–97. Most states' fiscal year begins on July 1 and ends on June 30.[7] Since January 1978, the Bureau of Labor Statistics has calculated a monthly unemployment rate for every state (expect California, first calculated in 1980). Since FY 1979, I calculate the fiscal year unemployment rate as the average monthly unemployment rate during the fiscal year. Prior to FY 1979, I calculate the fiscal year unemployment rate as a weighted average of the unemployment rates in the state in the two calendar years that comprise the fiscal years. The weights depend on the fraction of months for which the fiscal and calendar years overlap.

While the national business cycle is usually discussed in terms of changes in gross domestic product (GDP), the unemployment rate is a better measure of economic conditions in the state than gross state product (GSP). There are problems concerning the accuracy of GSP numbers. GSP is gross output minus the value of intermediate inputs. Evaluating the worth of intermediate inputs for the same company across different states is surely a daunting task. While such transfer pricing issues also arise for GDP, linkages across nations are both weaker and more carefully monitored than those across states. The final advantage of the unemployment rate is that during most of the period of study, it was measured monthly. This allows me to calculate a measure that corresponds in timing to the state financial year. By contrast, GSP is measured only yearly and is therefore more difficult to match accurately with the financial data. However, if I were to use the percentage change in GSP per capita as the measure of state fiscal condition instead of the fiscal year unemployment, I would arrive at a set of results broadly similar to those discussed below.[8]

Fiscal data

The data I use to measure state financial variables come from the annual survey of state government finances conducted by the U.S. Census Bureau (U.S. Department of Commerce, Bureau of the Census, 1977–87 and 1988–97). The survey measures approximately 450 different aspects of state revenues, expenditures, debts, and assets. I use the survey data from 1977–97; the 1998 data have not yet been released and the data prior to 1977 are not available in electronic form. Importantly, this is not accounting data drawn from state budgets, but is statistical in nature. Budgetary data would not be as comparable across states or over time. The variables measured over this period have been relatively consistent. One important exception is that major changes in measurement of debt occurred in 1988. (Throughout, dollar numbers are in GDP-deflated 1997 dollars.)

Methodology

In analyzing state fiscal behavior, I look at how a change in the fiscal year unemployment rate changes a variable measuring a fiscal outcome. I measure all fiscal outcomes in per capita terms to make the numbers comparable across states. Throughout, the unit of analysis is an individual state and states are not weighted in terms of population. I look at how a 1 percentage point change in the fiscal year unemployment rate (say, a jump from 4.2 percent to 5.2 percent) affects the per capita measure of a fiscal variable. Throughout the remainder of the analysis, I omit the state of Alaska. Alaska's fiscal behavior differs drastically from that of the other 49 states, mainly due to the revenues Alaska receives from oil production.

I also include a series of state fixed effects. This allows the average value of a variable to differ across states. This is especially important when looking at state expenditure patterns because the role of the local governments in service provision differs quite dramatically across states. Importantly, I do not include any measures of the nature or severity of state balanced budget requirements. One might want to include these interacted with the unemployment rate

to investigate whether fiscal variables in states with stricter requirements are more responsive to changes in the unemployment than states with more lax requirements; however, I do not do so here. I believe that the issuance of debt by all states implies that their provisions are more similar than different.[9] I am more interested in how all states behave because states as a whole are perceived as being more fiscally responsible than the federal government. I also include both a linear and a quadratic time trend to account for the fact that there was an upward secular trend in state spending during this entire period.

The regression estimated for each fiscal variable is:

$$\frac{\textit{fiscal variable}_{st}}{\textit{population}_{st}} = \alpha + \beta \times \textit{unemployment rate}_{st} + \chi \times (\text{year} - 1977)_t + \delta \times (\text{year} - 1977)_t^2 + \phi \times \textit{state dummies}_s + \varepsilon.$$

In the tables, I only present the coefficient on the unemployment rate, β. This coefficient can be interpreted as the effect of a 1 percentage point increase in the unemployment rate on the per capita amount of the fiscal variable. Note that the typical peak to trough difference in the unemployment rate is greater than 1 percent. For example, the average fiscal year state unemployment rate rose from 6.0 percent in 1981 to 9.8 percent in 1983. In the milder 1991 recession, the average fiscal year state unemployment rate increased from 5.2 percent in 1990 to 6.7 percent in 1992; it retreated to 5.0 percent in FY 1997. In some places I compare the behavior of the states to the behavior of the federal government. To do so, I use federal data from the Budget of the United States (U.S. Government, Office of Management and Budget, 1999). This is accounting data, unlike the state data. In the case of the federal data, I estimate the same regression presented above, excluding the series of state dummies.

I break the analysis into four parts—first, I look at the gap between state expenditures and revenues (the deficit or surplus); second, at state revenues; third, at expenditures; and finally, at state indebtedness and asset accumulation. In each section, I look separately at finances inside and outside the insurance trust funds run by the states. The states administer a number of different insurance trust systems, including employee retirement systems, unemployment compensation, workers' compensation, and other smaller funds (including disability and sickness policies). The budget items outside the insurance trust system are considered "general" budget items.[10]

Responsiveness of the surplus to the business cycle

Between 1977 and 1997, average state general fund revenues exceeded average state general fund expenditures by almost $64 per capita while average state insurance trust fund revenues exceeded average state insurance trust fund expenditures by nearly $189 per capita (see table 1, column 1). While these calculations imply that states operate with a general fund surplus on average, this is somewhat misleading because state expenditure data exclude state payments into their insurance trust systems. State contributions to their insurance trust systems average just over $70 per capita yearly. These contributions are almost

TABLE 1

Per capita budget deficit or surplus, 1977–97
(dollars)

Budget category	Average per capita value	Effect of 1 percentage point increase in unemployment rate
Total surplus (revenues – expenditures)	252.00	–23.03 (6.943)
General fund surplus	63.50	–10.85 (4.642)
Insurance trust surplus	188.50	–12.18 (5.822)
General fund surplus net of interest payments	163.87	–10.92 (4.526)

Notes: Absolute t-statistics are in parentheses. The final column of each row represents the coefficient on the unemployment rate in a separate regression. Other variables included in all regressions are a linear and quadratic time trend, a constant, and a series of state fixed effects.
Source: Author's calculations from U.S. Department of Commerce, Bureau of the Census, 1977–87 and 1988–99, *State Government Finances*.

exclusively payments by states into their employee retirement systems. If these intragovernmental transfers were included as general fund expenditures and insurance trust revenues, the average general surplus would become slightly negative and the average insurance trust surplus would increase.

When I run the regression specified above to look at how state surpluses are affected by changes in the unemployment rate, I find that a 1 percentage point increase in the unemployment rate decreases state surpluses by $23.03 per capita,[11] as shown in the last column of table 1. This combines a $10.85 ($2.34)—number in parentheses indicates the standard error—per capita drop in the general fund surplus with a $12.18 ($2.09) per capita drop in the insurance trust surplus. This result suggests that state budgets as a whole do respond to the business cycle. Below, I investigate the sources of this business cycle variation by exploring revenues and expenditures separately.

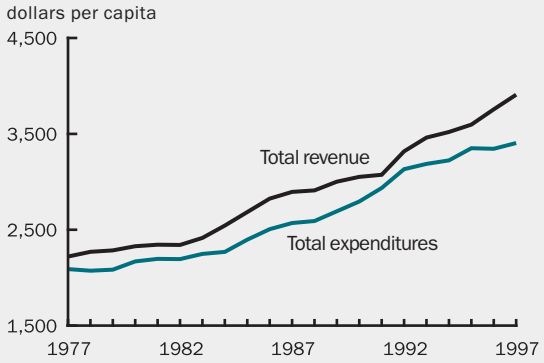
Responsiveness of revenues to business cycle

Between 1977 and 1997 average state yearly revenues per capita were $2,893. This breaks down into $2,448 raised by the general fund and $445 raised by the insurance trust systems. Total revenues per capita were growing rather steadily over the period, from $2,220 in 1977 to $3,908 in 1997 (see figure 1). These revenues come from five distinct sources: taxes, intergovernmental transfers from both the federal government and local governments, government charges for service provision,[12] funds from miscellaneous other sources including lotteries and property sales, and contributions to the trust systems run by the state. Table 2 presents both totals and the breakdown of average yearly per capita revenues during this period and the responsiveness of budget items to the unemployment rate. Figure 2 depicts the percentage contribution to total revenues from each of these sources. The table and figure demonstrate that the great majority of state government funds come from taxes, intergovernmental transfers from the federal government, and insurance trust contributions.

Overall per capita revenues are somewhat responsive to changes in the fiscal condition in the state as measured by the state fiscal year unemployment rate. In particular, as presented in table 2, I find that a 1 percentage point increase in the state unemployment rate decreases total state revenues by $13.80 ($4.16) per capita. This combines a $20.08 ($3.47) decrease in general revenues with a $6.28 ($2.12) increase in the revenues of the insurance trust funds. The changes mask considerable variation within the various categories in the budget.

FIGURE 1

Expenditures and revenues, 1977–97



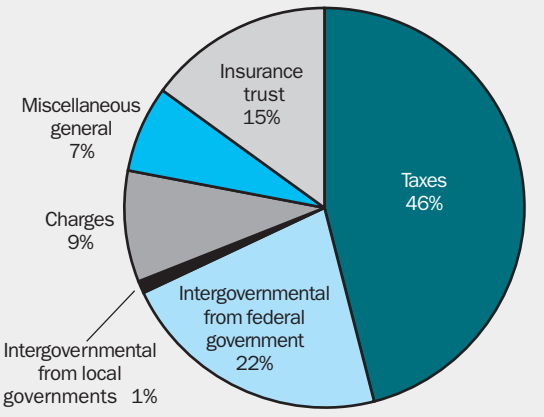
Note: Dollars are measured in 1997 dollars.
Source: Author's calculations from U.S. Department of Commerce, Bureau of the Census, 1977–87, *Survey of State Government Finances*, Washington, DC: Government Printing Office, and U.S. Department of Commerce, Bureau of the Census, 1988–97, *Survey of State Government Finances*, available on the Internet at www.census.gov/govs/state/ www.census.gov/govs/state/.

Taxes

Not surprisingly, taxes are among the most fiscally sensitive of state revenue sources. Although the lion's share of such revenues comes from sales and income taxes, state governments also assess license taxes and taxes on miscellaneous items such as stock transfers. Table 2 shows the breakdown in per capita tax revenues into these three categories and their responsiveness to a 1 percentage point change in the

FIGURE 2

State revenue sources, 1977–97



Source: Author's calculations from U.S. Department of Commerce, Bureau of the Census, 1977–87, *Survey of State Government Finances*, Washington, DC: Government Printing Office, and U.S. Department of Commerce, Bureau of the Census, 1988–97, *Survey of State Government Finances*, available on the Internet at www.census.gov/govs/state/www.census.gov/govs/state/.

TABLE 2

Average yearly revenue per capita, 1977–97 (dollars)

Budget category	Average per capita value	Effect of 1 percentage point increase in unemployment rate
Total revenues	2,892.54	–13.80 (3.313)
General fund revenues	2,447.97	–20.08 (5.788)
Tax revenues	1,318.45	–21.04 (8.728)
Sales taxes	662.80	–10.90 (7.524)
Income taxes	460.83	–10.50 (7.331)
Other taxes	194.82	0.36 (0.270)
Intergovernmental revenues	649.95	3.24 (2.040)
Federal intergovernmental revenues	625.70	2.74 (1.850)
Public welfare	276.85	4.51 (4.047)
Education	111.96	–1.27 (4.707)
Other	236.88	–0.50 (0.513)
State intergovernmental revenues	24.26	0.50 (1.409)
Charges	261.09	–2.01 (2.898)
Miscellaneous general revenues	218.47	–0.26 (0.188)
Insurance trust	444.57	6.28 (2.958)
Contributions	223.68	0.55 (0.669)
Investment revenue	208.22	–2.73 (1.390)
Federal unemployment insurance advances	12.67	8.46(12.479)

Notes and source: See table 1.

unemployment rate.[13] Some tax revenues are more sensitive to the business cycle than others. As table 2 indicates, sales and income tax receipts are far more sensitive to the business cycle than other taxes.

While I find that income and sales taxes are equally sensitive to the business cycle, I would expect income taxes to be far more sensitive. This expectation arises from the fact that while income is highly sensitive to the unemployment rate, individuals dip into savings in order to smooth consumption during downturns. As a result of this smoothing, total sales, and hence sales tax receipts, are not thought to be as sensitive as income taxes to the business cycle.

The lower than expected income tax numbers can be explained by the fact that these numbers represent the change in actual tax collections and do not account for the fact that states often make statutory changes in their tax structures to counteract the effects of the business cycle. In particular, states tend to raise tax rates during times of economic difficulty and lower taxes in times of economic strength. In the absence of such statutory changes, the cyclicality of state revenues would be more pronounced.[14] One potential explanation for the income tax number not being larger than the sales tax number is that income tax levels are more often statutorily adjusted than sales tax levels in response to economic conditions. This conjecture certainly holds true of the current economic expansion. In their yearly reports on *State Tax Actions* from 1995 to 1998, the National Conference of State Legislatures (NCSL) reported that income tax reductions and, in particular, reductions in the personal income tax "dominated state tax reduction efforts" (NCSL, 1995); were "the primary focus of state tax cuts" (NCSL, 1996); "dominated legislative tax actions" (NCSL, 1997); and were "the main focus of cuts" (NCSL, 1998). In contrast, in most years excise and sales tax changes were relatively small. In total, the tax reductions put into effect between 1995 and 1998 reduced state taxes by a staggering $16.8 billion dollars.

Even though states counteract some of the effects of the business cycle on tax receipts by changing tax rates, states are still faced with declining resources during times of economic difficulty. The tax rate changes do not totally counteract the fiscal effects of recession.

Intergovernmental revenues

Intergovernmental transfers are the second major source of state revenue. While states receive payments from both the federal and local governments, the amount from the federal government far exceeds the amount from the local governments (see table 2). As shown in table 2, intergovernmental revenues are relatively unresponsive to business cycle conditions. Looking at the breakdown into local and federal intergovernmental revenues yields a similar picture—in both categories per capita revenues increase slightly when the unemployment rate increases.

To look at the relationship between federal intergovernmental revenues and the business cycle a bit more closely, I break revenues into three categories—education, public welfare and other. Public welfare consists of grants for income support and medical assistance programs. I expect intergovernmental spending on public welfare revenues to be more cyclically sensitive than spending in the other categories. The results in table 2 support this picture. Intergovernmental revenues for public welfare increase when the economy worsens, while spending in the other two categories declines. Importantly, the welfare reform legislation passed in 1996 will reduce the cyclicality of public welfare grants because it replaced an open-ended matching grant with a fixed block grant.[15]

Charges

Charges include government fees for service provision and revenues from the sale of products in connection with general government activities. For example, the air transportation measure of charges includes landing fees at airports and rents for concession stands. I also include the revenues of public utilities and liquor stores in this category. As is shown in table 2, revenues from charges only decline slightly during a downturn.

Miscellaneous revenue sources

Miscellaneous revenue sources consist of monies coming into the state that cannot be easily categorized elsewhere. These include proceeds from special assessments and property sales and monies from interest earnings, rents, royalties, fines, forfeits, and state lotteries. The analysis of miscellaneous revenues differs from that of other revenue sources because a major code change in FY 1988 makes a couple of the subcategories noncomparable before and after this date. Since 1988, a 1 percentage point change in the unemployment rate has increased miscellaneous revenues by $4.82 ($2.06), while prior to 1988, a 1 percentage point change in the unemployment rate decreased revenues by $3.39 ($1.88). (I present the regressions for the entire period in table 2 so that the subcategories can add up to the total). The more recent experience suggests that state governments can expect revenues to go up slightly in the future when the economy worsens.

One argument regarding how the federal government might adjust its budgeting in order to achieve budget balance in times of economic stress is that it might engage in "increased sales of public lands" (Eisner et al., 1997). I explore whether the state governments engage in the analogous activity by increasing property sales during times of high unemployment. Because the category "property sales" did not experience a definitional change in 1988, I look at behavior over the entire sample period.[16] I find no evidence of increased property sales during times of economic stress. While this does not mean that the federal government, with its far more extensive land holdings, would not engage in this behavior, it does suggest that states do not sell property to compensate for budget shortfalls.[17]

Insurance trust revenues

Revenues for insurance trust programs come from three different sources (aside from within the state itself): contributions from employees, contributions from other governments (both local and federal), and interest revenues.[18] As shown in table 2, overall insurance trust revenues are countercyclical, increasing $6.28 ($2.13) when the unemployment rate increases by 1 percentage point.

Not surprisingly, most of the variation within this category over the business cycle occurs in unemployment compensation. In particular, federal advance contributions, the amounts credited to the states when contributions and interest cannot pay unemployment benefits due, increase by $8.46 ($0.68) per capita when the unemployment rate increases by 1 percentage point. By contrast, contributions and investment revenues are much less sensitive to the state of the economy.

Revenue results and implications

During times of economic difficulty, state revenues drop by about $23 per capita. This drop is mostly driven by declining tax revenues and in particular by declining income and sales tax receipts. There are three principal reasons that this decline is not more pronounced. First, state income tax rates are often increased when times are bad. Although this does not emerge directly from this analysis, the recent declines in state tax rates highlight this phenomenon. Second, the states get more money from the federal government during downturns, particularly in terms

of intergovernmental funds for public welfare and federal advances from the unemployment insurance system. Third, state governments rely on a number of income sources that are fairly acyclical. Only 44 percent of state revenues come from taxes and only 15 percent come from the highly sensitive income tax. By contrast, 53 percent of federal government revenues came from taxes in 1991 and 47 percent came from income taxes (U.S. Department of Commerce, Bureau of the Census, 1994).

While state revenues decline in recessions, federal government revenues have historically declined even more. Between 1977 and 1997, a 1 percentage point increase in the national unemployment rate reduced federal government revenues per capita by $115.75 ($30.00), 2.5 percent of the mean federal revenue level of $4,674.06; by contrast the drop in state revenues is about 0.8 percent of mean revenues ($23.03 of $2,892.54).

The methods that states use to mitigate this decline, heavier reliance on the federal government, tax increases, and use of less cyclically sensitive revenue sources, would not be as readily available to the federal government. Heavier reliance on a higher level of government is obviously not an option for the federal government. Tax increases during downturns are a possibility but would aggravate recessions by decreasing disposable income and consumption during recessions. States are able to increase tax rates because they are not responsible for the condition of the macroeconomy. Eventually the federal government may want to seek out less cyclically sensitive revenue sources. One such possibility would be a national sales tax that could be less sensitive than the income tax to downturns.

Because of the super-majority requirement for revenue increases enshrined in most balanced budget proposals, it is unlikely that much of the adjustment in recessions would occur via revenues. Indeed, this is exactly the point for some proponents of the measure—they seek an amendment that would force Congress to cut spending during difficult times. Next, I investigate what happens to state expenditures during recessions.

Responsiveness of expenditures to business cycle

State government expenditure is divided into five different categories—current spending, capital spending, intergovernmental expenditures, interest on the debt, and insurance trust expenditures. The breakdown of expenditures is presented in the first column of table 3 and in figure 3. Like revenues, state per capita expenditures have been steadily increasing since 1977 (see figure 1).

Overall expenditures are somewhat sensitive to business cycle conditions, although less so than revenues. The first row of table 3 shows that a 1 percentage point increase in the unemployment rate increases overall expenditures by $9.23 ($4.14) per capita. This is the combination of a $9.23 ($3.75) decline in general fund expenditures with an $18.46 ($0.95) increase in insurance trust expenditures. Falling general fund expenditures are more than offset by rising insurance trust spending.

Current expenditure

Current expenditure represents the biggest portion of state government expenditure at just over half of the entire category. Current operations include spending on a vast array of goods and services including transportation, hospitals, state educational institutions, and public welfare.[19] As shown in table 3, current expenditure is rather flat over the business cycle, increasing by an insignificant amount when the unemployment rate rises.

Breaking current operations expenditures down by the function they support, I find that during downturns public welfare spending increases, while spending on education (mostly higher education) and other services falls. The increase in public welfare is not surprising given that during downturns a greater fraction of the population relies on the government for support.

Capital expenditure

Capital expenditure is much more sensitive to the business cycle than current expenditure. Table 3 shows that capital expenditure per capita drops by $6.94 ($1.23) when the unemployment rate increases by 1 percentage point. This drop is evenly split between a decline in spending on construction and a decline in other capital outlay (mostly comprising land and equipment).[20]

Given that the benefits of capital projects are less immediately apparent, spending on capital projects may be politically easier to cut. In addition, states have more discretion over capital spending because it is less likely than current spending to arise from entitlement programs. Capital spending is also naturally less persistent. Although a state cannot easily close a university to bring about budget balance, it can slow down major capital projects or wait to begin new ones.

The role of this reduction in capital spending is interesting in light of the fact that state capital budgets are outside the operating budgets directly affected by balanced budget restrictions. It suggests that

TABLE 3

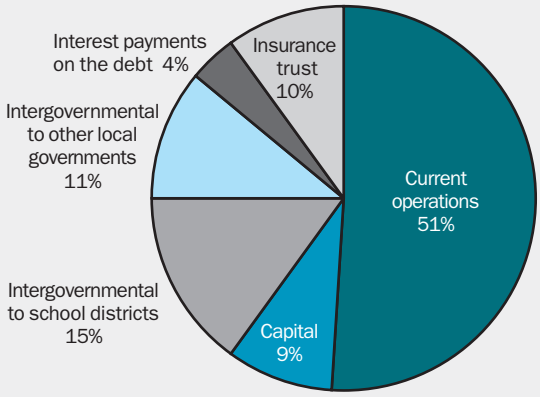
Average yearly expenditure per capita, 1977–97
(dollars)

Budget category	Average per capita value	Effect of 1 percentage point increase in unemployment rate
Total expenditures	2,640.54	9.23 (2.232)
General fund expenditures	2,384.47	–9.23 (2.462)
Current operations	1,349.71	2.75 (1.102)
Education	358.64	–2.43 (4.190)
Public welfare	404.66	7.09 (4.776)
Other current operations	586.41	–1.91 (1.288)
Capital expenditure	239.85	–6.94 (5.640)
Construction	192.98	–3.57 (3.398)
Other capital outlay	46.88	–3.37 (7.996)
Intergovernmental expenditures	694.54	–4.97 (3.180)
To school districts	386.76	–4.22 (3.158)
To other local	302.19	–0.77 (0.648)
To federal government	5.59	0.02 (0.294)
To education	464.57	–3.74 (3.077)
To public welfare	43.42	2.00 (4.254)
To other	186.55	–3.23 (4.248)
Interest payments on the debt	100.37	–0.07 (0.118)
Insurance trust expenditure	256.07	18.46 (19.500)
Unemployment benefits	99.23	17.71 (28.736)
Other trust payments	156.84	0.75 (1.039)

Notes and source: See table 1.

FIGURE 3

State expenditure areas



Source: Author's calculations from U.S. Department of Commerce, Bureau of the Census, 1977–87, *Survey of State Government Finances*, Washington, DC: Government Printing Office, and U.S. Department of Commerce, Bureau of the Census, 1988–97, *Survey of State Government Finances*, available on the Internet at www.census.gov/govs/state/ www.census.gov/govs/state/.

states reduce pressure on their operating budget by reducing capital spending. When I compare debt issuance to capital spending, I find that if all revenues from debt issuance were spent on capital projects, only 60 percent of the money for capital projects would be financed by debt.[21] This indicates that states finance a large portion of capital expenditure out of current revenues.

Intergovernmental expenditure

States transfer money to local governments and to the federal government. The great majority of these funds go to school districts and to general-purpose local governments, such as county, municipal, and township governments. Only a small sum is transferred to the federal government. As shown in table 3, overall intergovernmental expenditures fall when the economy worsens.

I break up intergovernmental expenditures in two different ways. First, I divide them by recipient government: school districts, other local governments, and federal government. Second, I divide them by

function: education, public welfare, and other. While transfers to the federal government and to local governments are relatively flat over the business cycle, transfers to school districts drop off significantly when the economy worsens. The functional breakdown yields the same picture, with declines in education spending being the main explanation for the overall reduction in intergovernmental revenues. By contrast, as with federal intergovernmental revenues and current operations, public welfare intergovernmental spending increases during downturns as states transfer more money to localities to support swelling public assistance rolls.

Interest expenditures

States pay interest on general debt and interest on the debts of public utilities, with the general debt accounting for the bulk of interest paid. As shown in table 3, interest expenditures are largely acyclical. Although state debt may increase during difficult economic times, as explained further below, the stock of debt and, hence, interest payments are quite flat over time.

Insurance trust

Insurance trust expenditures are benefit payments to recipients under the state's employee retirement, workers compensation, unemployment insurance, and other trust funds. In total, as shown in table 3, insurance trust expenditures are highly procyclical, increasing by $18.46 ($0.95) or about 7 percent of the mean when the unemployment rate increases by 1 percentage point.

Given that unemployment benefits are one source of insurance trust expenditures, the size of this increase is not surprising. During times of high unemployment, unemployment benefit benefits greatly increase. In fact, all of the increase in insurance trust spending that occurs when unemployment is high can be attributed to increases in spending for unemployment benefits.

Expenditure results and implications

During times of economic difficulty, states are able to decrease their general fund expenditures by $9 per capita in spite of increasing pressure on public welfare spending. States do this in three ways: They decrease higher education current expenditure; they drastically reduce capital expenditure; and they cut the funds going to school districts.

The implications of this for the federal government are mixed. There is no reason to believe that the federal government would not be able to cut current expenditure in some areas in response to recessions. While the size of federal government entitlement programs limits government flexibility, the federal government has some areas of responsibility that are akin to state governments' higher education responsibilities. The most obvious area is that of education, training, employment, and social services, but cuts in other areas would also be possible.

The states' ability to decrease capital spending is important in helping them to achieve budget balance. In fact, the drop in state capital spending almost totally offsets the increase in current public welfare expenditure brought about by a 1 percentage point increase in the unemployment rate. However, whether the federal budget would or should follow the states' lead in this arena is a difficult question. Some of the same factors causing the states to decrease capital spending during recessions may also affect the federal government. In particular, because capital spending has current costs and longer term benefits, cuts in capital spending may be politically easier to swallow than cuts in federal spending on education or job training. In addition, the absence of a federal capital budget may make federal capital spending even more responsive to economic conditions. It is possible that states do not reduce their capital spending further because they can issue debt for capital projects. Therefore, their ability to alleviate general budget pressures is limited by the portion of capital spending that is being financed by current revenues.

However, there is one important reason that federal capital spending may not be as susceptible to the business cycle as state capital spending. While the majority of state capital spending is for highways and higher education, projects that may be easy to delay, the great majority of federal capital expenditure goes to finance defense. Between 1977 and 1997, 82 percent of the money spent on direct federal capital expenditure was used for defense.[22] In no year did defense spending drop below 70 percent of total direct capital expenditure. It seems unlikely that federal defense spending would or should be a function of business cycle conditions. A brief glance at the numbers demonstrates that, historically, defense capital spending has been more a function of the political climate and whether the nation is at war than of the unemployment situation. For example, from 1943–46, at the height of U.S. involvement in World War II, defense capital goods represented about 99 percent of federal capital expenditure on average. The federal government could cut capital spending in other areas, but nondefense capital spending is a very small part of the budget—averaging only 1.6 percent between 1977 and 1997 (total capital spending averaged 9 percent of the federal budget over the same period).

In addition to reducing current spending for education and capital expenditure, state governments reduce overall intergovernmental grants, especially those to school districts. In general the states take advantage of their unique position in the intergovernmental structure by procuring additional grants from the federal government while sending less money to the local governments. The federal government could follow the states lead here by reducing intergovernmental expenditures to the states during times of economic stress. While this may improve the federal government's budgeting position, it would make it more difficult for the states to balance their budgets. Part of the reason state governments are able to come close to balancing their budgets is that the federal government does not achieve a balanced budget.

The federal government could not avail of the overall expenditure strategy relied on in the states because of its unique responsibility to provide for national defense. By contrast, the federal government may be able to follow the states' lead in cutting current expenditure and in cutting grants to lower levels of government. The wording of the federal balanced budget amendment implies that the government would need to cut spending to compensate for the entire drop in revenues. However, state governments have an important safety valve in their ability to issue debt to fund capital projects. Next, I investigate the extent to which they take advantage of this safety valve.

What happens to debt and assets?

The combination of the revenue and expenditure pictures for both the general and insurance trust funds is not very consistent with the common notion of budget balance. During difficult times, general fund revenues fall more than expenditures, and trust fund expenditures increase more than revenues. This implies that states must either deplete assets or issue debt when the economy deteriorates. In other words, their net asset position must worsen. Below, I look at what happens to state debt issuance and state reserve funds, both inside and outside the insurance trust system.

Short-term debt

Short-term debt is issued to account for unexpected shortfalls. This category includes debt payable within one year of issuance or debt backed by taxes to be collected in the same year. It includes items such as tax anticipation notes and short-term warrants and obligations, but excludes accounts payable and similar less formal non-interest-bearing obligations. States that are not allowed to carry over deficits still sometimes have short-term debt in the form of tax obligation notes and similar liabilities.

The Census Bureau only collects two short-term debt items (in stark contrast to the approximately 50 different measures of long-term debt)—the amounts outstanding at the beginning and the end of the fiscal year. I use the amount outstanding at the end of the year; given that most short-term debt has a maturity of under one year, this is a reasonably good proxy for issuance.[23] As table 4 shows, short-term debt is fairly responsive to the business cycle, increasing by about $2.41 ($0.56) for a 1 percentage point increase in the unemployment rate. However, this only goes part of the way in explaining how states finance the growing gap between revenues and expenditures during downturns. States also rely on additional long-term debt issuance.

Long-term debt and government assets

Because long-term debt and asset data before and after 1988 are not comparable (due to a classification change in 1988), I use data from 1989 onwards. State government long-term debt and asset data are far more complicated than other financial data for three main reasons. First, over 40 percent of state government debt is "public debt for private purposes." This debt is issued using the tax-exempt status of state governments to finance expenditures by private firms. I analyze this debt separately from government purpose debt.[24] Second, not all debt issuance funds contemporaneous expenditures. Some debt is issued to refund previously issued debt. Because a lot of state debt is callable (that is, it can be redeemed prior to maturity for a prespecified premium), when interest rates are falling, states can realize savings if they call debt and refund it at a lower interest rate. Because I am interested in debt issuance that contributes to the state's concurrent fiscal situation, I would ideally like to look only at new government purpose debt issued, that is, net of refunding. Unfortunately, I cannot do this because debt issued for refunding cannot be separated into public and private purpose debt. Third, an analysis of debt cannot be separated from an analysis of government assets because two of the three state government asset measures are directly related to debt. The sinking fund contains money explicitly saved for debt redemption, while the bond fund contains the proceeds of bond issuance prior to disbursement. Only the "other funds" category contains assets not explicitly linked to debt. Because these assets are all stocks rather than flows, I look at the change in value per capita from one year to the next as the appropriate measure of government assets.[25]

TABLE 4

Debts and assets, 1989–97
(dollars)

Budget category	Average per capita value	Effect of 1 percentage point increase in unemployment rate
Short-term debt, 1977–97		
Outstanding at end	14.94	2.41 (4.299)
Long-term debt, 1989–97		
Issuance		
Governmental purposes	155.86	16.06 (2.604)
Private purposes	145.37	10.90 (1.755)
Refunding	37.50	10.89 (2.680)
Redemption/retirement		
Governmental purposes	104.77	18.99 (4.228)
Private purposes	109.64	8.26 (1.815)
Retired by refunding	36.11	9.93 (2.619)
Government assets, 1989–97		
Change in sinking fund	–8.10	–10.97 (2.553)
Change in bond fund	0.71	–7.51 (2.432)
Change in other funds	35.83	3.61 (0.489)
Insurance trust assets, 1978–97		
Change in employee retirement	183.01	2.84 (0.466)
Change in unemployment insurance	7.43	–8.54 (7.730)
Change in worker's compensation	8.35	–0.33 (0.268)
Change in other trust assets	0.11	0.04 (0.524)

Notes and source: See table 1.

Table 4 shows the relationship between the state unemployment rate and the state long-term debt issuance, redemption, and asset measures. The first thing to notice is that all measures of debt issuance increase significantly during downturns. Because nearly all long-term debt is used to finance capital projects and because capital spending drops off quite significantly during downturns, the increase in debt issuance suggests that state governments finance a higher percent of their capital spending with debt during recessions. This implies that states use debt issuance as an important safety valve during recessions. The decrease in the state bond fund, also shown in table 4, supports this finding. Although states spend less on capital projects, they both draw down unspent monies from previous bond issuance and issue more bonds.

As with issuance, all three measures of debt redemption also increase during downturns (also in table 4). This result is more intuitive than it may appear when combined with the information on the change in the value of the sinking fund.[26] States redeem more debt during downturns, but it appears that this extra money is coming from a combination of debt refunding (which increases by $9.93 per capita) and a drop in the value of the sinking fund (which decreases by $10.97 per capita) rather than from current revenue sources. The transfers from the sinking fund probably occur because of cyclical changes in financial market conditions. In particular, states have an incentive to pay off debt using sinking fund assets when they are paying more interest on existing debt than they are receiving from fund assets. In short, during good times, states accumulate assets in their sinking funds that are then spent to call bonds when the economy worsens and interest rates fall. Finally, there is no evidence of changes in the assets of non-bond-related funds.

Assets of the trust funds

One of the most frequently articulated worries about a balanced budget requirement is that it would lead to the depleting of social security reserves in a downturn. Do state government deplete the assets of state managed trust funds in downturns? I look at the change in the assets of all four types of government trust funds—employee retirement, workers

compensation, unemployment insurance, and others. The employee retirement trust fund is the only one that is directly comparable to social security. The other funds, particularly the unemployment insurance trust fund, are *supposed* to fall during recessions.

Table 4 shows that there is little evidence of systematic raiding of the trust funds. While state unemployment insurance trust funds decline dramatically during downturns, there is no evidence that the assets of other trust funds fall.

Debt and assets results

I find that states issue more short-term and long-term debt during recessions. As mentioned above, the federal balanced budget amendment does not allow any new debt issuance short of a super-majority vote. Therefore, this avenue would not be open to the federal government. Instead, the federal government would be compelled to find areas in which to cut spending in order to confront revenue declines.

Conclusion

Both state and national balanced budget supporters frequently cite the experience of the states to demonstrate the feasibility of a federal balanced budget amendment. State governors and U.S. presidents alike have claimed that the state experience is a relevant example to the federal government. In this analysis of the way that state budgets respond to the business cycle, I find few examples of methods for budget balance used by the states that are directly relevant to the federal government. This is the case for four principal reasons.

First, state balanced budget requirements differ in one major way from the amendment currently contemplated at the federal level. State governments can and do issue both short-term and long-term debt to finance shortfalls and capital projects, respectively. The states are able to issue long-term debt because state capital projects are outside the restrictions imposed by the balanced budget amendments.

Second, despite the fact that state capital budgets are separate, states cut capital spending quite drastically during downturns in order to relax budgetary pressures. The current costs and delayed benefits of capital spending make it politically easier to cut. The federal government may not find capital spending so easy a target because most federal capital spending is for defense.

Third, states take advantage of their unique position in the federal system to cut funds going to local governments while drawing on increased funds from the federal government. The federal government can not draw down more money from a higher level of government, but could potentially decrease the money it sends to the states.

Finally, states increase tax rates during downturns and decrease them during booms. The states are able to engage in this behavior because, unlike the federal government, they are not perceived as being responsible for the macroeconomy.

Overall, the state experience with budget balance and business cycles is not a very relevant model for the federal government. State governors are not responsible for the macroeconomy or for national defense and, in general, confront a more relaxed budget restriction than that proposed for the federal government. Policymakers need to consider carefully how budget balance at the federal level could be achieved during an economic downturn under a balanced budget restriction—for example, which taxes could be increased, which programs could be cut, or which assets could be sold.

NOTES

[1]Briffault (1996) provides an interesting set of quotations suggesting that the state experience is relevant to the federal government.

[2]The National Association of State Budget Officers (1997) states that 40 of 48 states that responded to a survey report that their capital planning occurs in a capital budget.

[3]The exact definition of what capital spending consists of differs by state. This is the most common definition.

[4]Forty-eight states have either a constitutional or statutory balanced budget requirement. One state that does not is not permitted to carry over deficits. These combine to generate the frequently cited figure that 49 states have balanced budget restrictions. The exception is Vermont.

[5]For a further discussion of limits on long-term debt, see McGranahan (1999b).

[6]These comparisons actually underestimate the difference between the states and the federal government because, while the federal numbers are net of trust fund interest revenues, the state numbers are gross. I do not net out state interest revenues because the definition of interest revenues changed in 1988 to include revenues from public debt for private purposes. Therefore, it is impossible to calculate a net number for the states that is consistent over time. The gross numbers for the federal government would be 18 percent for expenditures and 21 percent for revenues.

[7]The year refers to the calendar year in which the fiscal year ends, so fiscal 1999 ended in most states on June 30, 1999. Some states

have different fiscal years. I take these differences into account when calculating the fiscal year unemployment rate.

[8]One disadvantage of using the unemployment rate is that it is often viewed as a lagging indicator of economic activity.

[9]For a discussion of the effects of different balanced budget restrictions in the states, see Poterba (1994).

[10]This division is analogous to the separation between on-budget and off-budget in the federal context, because the federal budget excludes most social security funds.

[11]With a standard error of $3.32; note that table 1 shows t-statistics rather than standard errors.

[12]I include receipts of utilities and liquor stores run by the state in charges. In Census Bureau statistics, these are treated separately. They are generally very small and do not warrant separate treatment.

[13]Sales taxes refer to all sales and gross receipt taxes, including general sales, gas, and tobacco taxes. Income taxes refer to both individual and corporate income tax collections.

[14]For a discussion of tax revenue changes taking statutory changes into account, see Dye and McGuire (1998).

[15]For further discussion of this issue, see McGranahan (1999a).

[16]The classification manual defines property sales as "amounts received from sale of real property, buildings, improvements to them, land easements, rights-of-way, and other capital assets (buses, automobiles, etc.), including proceeds from sale of operating and nonoperating property of utilities. Includes sale of property to other governments."

[17]Interestingly, the historical relationship between federal property sales and the unemployment rate has been negative, indicating that the federal government sells less when the economy is bad.

[18]Contributions by the state to its own insurance trust systems are considered within government transfers and do not enter the revenue tabulations.

[19]I include spending on assistance and subsidies in the current expenditure category. It is only a small portion of total current expenditure. In published Census tables, assistance and subsidies (which include scholarships, veterans benefits, and some welfare payments) are usually presented separately.

[20]There were some minor changes in coding of some of the capital outlay variables in 1988. Looking only at data from after this change yields very similar conclusions—capital expenditure falls off, mostly driven by changes in spending on equipment and existing land and structures.

[21]The 60 percent number represents the average of debt issuance divided by capital spending from 1988–97. Debt issuance excludes debt for private purposes but is not net of refunding.

[22]Direct capital expenditure excludes grants. Many grants are to state governments for highways and other programs.

[23]The U.S. Department of Commerce, Bureau of the Census (1995) reports that "obligations having no fixed maturity date (even where outstanding for more than one year if payable from a tax levied for collection in the same year it was issued)" are included in short-term debt.

[24]The major reclassification in 1988 pertains to changes in the categorization of public debt for private purposes. Prior to 1988 it is not possible to fully separate it from other debts. The spending supported by public debt for private purposes does not show up in the states' expenditure measures.

[25]This is not the per capita change, but the change per capita where the population is the population in the second year.

[26]I subtract the value of public debt for private purposes outstanding from the sinking fund numbers to account for the fact that the value of collateral pledged for private purpose debt is included in the sinking fund numbers.

REFERENCES

Bond Market Association, The, 1999, "Daily report of municipal bond transactions" available on the Internet at www.investinginbonds.com/, accessed April 20.

Branstad, Terry E., 1997, "Balancing the budget: What Washington can learn from the states," Heritage Lecture, No. 586, available on the Internet at www.heritage.org/library/categories/budgettax/hl586.html, May 13.

Briffault, Richard, 1996, *Balancing Acts*: *The Reality Behind State Balanced Budget Requirements*, New York: The Twentieth Century Fund Report.

Dye, Richard F., and Therese J. McGuire, 1998, "Block grants and the sensitivity of state revenues to recession," in *National Tax Association Proceedings of the Annual Conference on Taxation*, Washington, DC, pp. 15–23.

Eckl, Corina, 1998, "States broaden the scope of rainy day funds," Washington, DC: National Association of State Legislatures, available on the Internet at www.ncsl.org/programs/fiscal/rdf97.htm.

Eisner, Robert, with Robert M. Solow, and James Tobin, 1997, "Petition in opposition to Balanced Budget Amendment," in "Economists Oppose the Balanced Budget Amendment," Max Sawicky (lead contributor), Internet discussion board of Communications for a Sustainable Future, available at http://csf.colorado.edu/mail/pkt/jan97/0338.html, January 16.

Engler, John, 1997, "Text of the State of the State address," *The Detroit News*, available on the Internet at http://detnews.com/1997/metro/9701/29/01290057.htm, January 29.

Illinois, State of, 1970, *Constitution of the State of Illinois*, adopted at special election on December 15, available on the Internet at www.legis.state.il.us/commission/lrb/conmain.htm.

Keating, Frank, 1997, "The State of the State," available on the Internet at www.oklaosf.state.ok.us/osfdocs/sos97.html, February 3.

McGranahan, Leslie, 1999a, "Welfare reform and state budgets," *Chicago Fed Letter*, Federal Reserve Bank of Chicago, No. 137, January.

_____, 1999b, "Voter preferences for capital and debt spending: Evidence from state debt referenda," *Proceedings of the Ninety First Annual Conference on Taxation*, Washington, DC: National Tax Association, forthcoming

National Association of State Budget Officers, 1997, "Capital budgeting in the states: Paths to success," available on the Internet at www.nasbo.org/pubs/capbud97/capbud97.htm/.

National Conference of State Legislatures (NCSL), 1998, *State Tax Actions 1998*, Washington, DC: National Conference of State Legislatures, available on the Internet at www.ncsl.org/programs/fiscal/sta97sum.htm, accessed April 1, 1999.

_____, 1997a, *Capital Budgeting in the States*, Washington, DC: National Association of State Budget Officers, September.

_____, 1997b, *State Tax Actions 1997*, Washington, DC: National Conference of State Legislatures, available on the Internet at www.ncsl.org/programs/fiscal/sta97sum.htm, accessed April 1, 1999.

_____, 1996, *State Tax Actions 1996*, Washington, DC: National Conference of State Legislatures, available on the Internet at www.ncsl.org/programs/fiscal/STAEX.HTM, accessed April 1, 1999.

_____, 1995, *State Tax Actions 1995*, Washington, DC: National Conference of State Legislatures, available on the Internet at www.ncsl.org/programs/fiscal/STA95P1.HTM, accessed April 1, 1999.

Poterba, J., 1994, "State responses to fiscal crises: The effects of budgetary institutions and politics," *Journal of Political Economy*, Vol. 102, No. 4, pp. 799–821.

Snell, Ronald K., 1999, "State balanced budget requirements: Provisions and practice," *National Conference of State Legislatures Fiscal Letter*, available on the Internet at www.ncsl.org/programs/fiscal/0796fl.htm.

U.S. Department of Commerce, Bureau of the Census, 1999, *Governments Finance and Employment Classification Manual*, Washington, DC, available on the Internet at www.census.gov/govs/www/class.html/.

_____, 1998, *1997 State Government Finance Tables by State*, available on the Internet at www.census.gov/govs/www/stsum97.html.

_____, 1994, *Government Finances: Summary of Federal Government Finances—1991 to 1994*, available on the Internet tat www.census.gov/govs/fedfin/federal.txt.

_____, 1990, *State Government Finances in 1989*, Washington, DC: U.S. Government Printing Office, Series GF-89-3.

_____, 1988–97, *State Government Finances*, available on the Internet at www.census.gov/govs/state/.

_____, 1977–87, *State Government Finances*, Washington, DC: U.S. Government Printing Office, electronic data provided by the Census Bureau Governments Division, Series GF-Year-3.

U.S. Government, Office of Management and Budget, 1999, "Budget of the United States government, fiscal year 2000, historical tables," Washington, DC: U.S. Government Printing Office, available on the Internet at www.access.gpo.gov/usbudget/fy2000/.

U.S. Senate, 1997, "S.J. Res. 1, proposing an amendment to the Constitution of the United States to require a balanced budget," available on the Internet at http://thomas.loc.gov/, accessed April 19, 1999.

Birth, growth, and life or death of newly chartered banks

Robert DeYoung

Introduction and summary

Thousands of new commercial banks have been chartered in the U.S. over the past two decades. As the U.S. banking industry continues to consolidate, these *de novo* banks are potentially important for preserving competition and providing credit in local markets. However, like other new business ventures, newly chartered banks initially struggle to earn profits, and this financial fragility makes them especially prone to failure. In this article, I document the financial evolution of the typical de novo bank and develop and test a simple theory of why and when new banks fail.

Recent decades have seen an upsurge in the number of mergers and failures among new banks. Figure 1, panel A shows the annual change in the number of commercial bank charters in the U.S. since 1966. Prior to 1980, the reduction in bank charters due to mergers and failures was relatively stable at about 100 charters per year, or about 1 percent of the industry total (figure 1, panel B). The pace accelerated greatly after 1980, and since 1986 about 600 charters, or 5 percent to 6 percent of the industry total, have disappeared each year due to mergers and failures.

To a large extent, this tremendous consolidation can be explained by the repeal of federal and state laws that restricted branch banking and interstate banking. As these restrictions gradually were relaxed, banking companies expanded their geographic reach by acquiring thousands of other banks, and reduced their overhead expenses by converting thousands of affiliate banks into branch offices. This geographic expansion, combined with newly deregulated deposit rates, increased competition between commercial banks just when new information technology was allowing mutual funds, insurance companies, and the commercial paper market to compete for banks' traditional loan and deposit businesses. Under these new competitive conditions, many commercial banks became more vulnerable to economic downturns, and thousands of banks failed during the 1980s and early 1990s. Over the past two decades, the combined effect of these mergers and failures has reduced the number of commercial banks in the U.S. by nearly 40 percent.

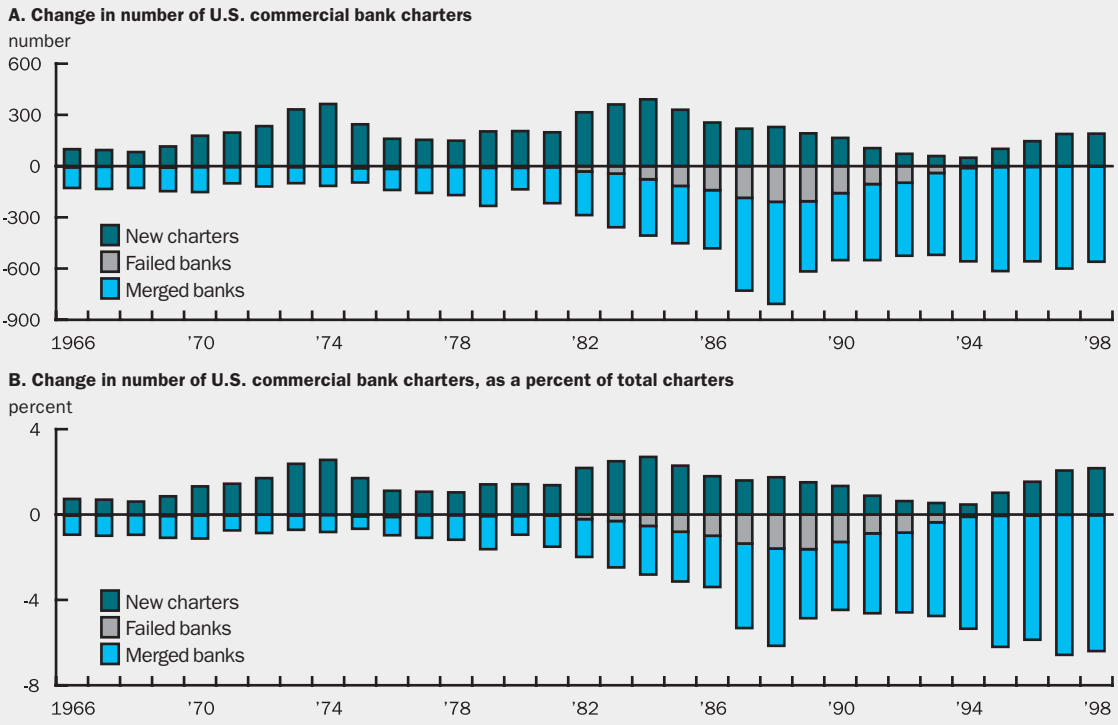
This consolidation has been partially offset by a recurring wave of new bank charters. As shown in figure 1, panel A, over 3,000 de novo commercial banks have been chartered by state and federal banking authorities since 1980. It is generally believed that these newly chartered banks can help restore competition in local markets that have experienced a large amount of consolidation. It is also commonly believed that these newly chartered banks can help replace credit relationships for small businesses whose banks failed or were acquired or reorganized. However, before a newly chartered bank can provide strong competition for established banks and before it can be a dependable source of credit for small businesses, *it must survive long enough to become financially viable*.

I begin by examining the conditions under which investors are likely to start up new banks, including the influence of business cycles, merger activity in local banking markets, and the policies of federal and state chartering authorities. Next, I track the evolution of profits, growth rates, capital levels, asset quality, overhead costs, and funding mix at more than 1,500 commercial banks chartered between 1980 and 1994. These data suggest that newly chartered banks pass through a period of financial fragility during which they are more vulnerable to failure than established

Robert DeYoung is a senior economist and economic advisor at the Federal Reserve Bank of Chicago. *The author thanks Iftekhar Hasan and Curt Hunter for their advice; Richard Cahill, Philip Jackson, and participants at a Federal Reserve Bank of Chicago seminar for helpful comments; Eli Brewer and David Marshall for reading an earlier draft of this article; and especially, Nancy Andrews for outstanding data support.*

FIGURE 1

Entry and exit of commercial banks



Source: Federal Deposit Insurance Corporation, 1966–98, "Change in number of insured commercial banks," available on the Internet at www2.fdic.gov/hsob/.

banks. Specifically, new bank capital ratios quickly decline to established bank levels, but new bank profits improve more slowly over time before attaining established bank levels.

Based on these empirical observations, I develop a simple *life-cycle* theory of de novo bank failure, in which the probability of failure at first rises, and then declines with the age of the new bank. I use *hazard function analysis* to test this simple theory for 303 new commercial banks chartered in 1985, just as the wave of bank failures shown in figure 1 was picking up steam. The tests offer support for the simple theory. On average, the results suggest that newly chartered banks are less likely to fail than established banks during the first few years of their lives; however, new banks quickly become substantially more likely to fail than established banks; and, over an extended period of time, new bank failure rates gradually converge to the failure rates of established banks.

What are the implications of these results for bank supervision and bank competition policy? The results suggest that the policies in place during the 1980s successfully insulated new banks from economic disruptions early in their lives, but were less successful in preventing new banks from failing after the initial years. Clearly, de novo bank failure rates could be reduced by requiring investors to supply higher amounts of start-up capital or by requiring banks to maintain extranormal capital-to-asset ratios in the early years—indeed, the latter policy option was adopted by federal bank supervisors during the 1990s. However, *failure-proofing* de novo banks is not an optimal policy. The social costs of small bank failure are relatively low, and setting higher capital requirements would at some point discourage investment in new banks and thereby limit the competitive benefits of de novo entry.

Birth of new banks

As illustrated in figure 1, panel A, the number of new banks started up each year has ebbed and flowed over the past three decades. There are a number of explanations for these patterns. Like all new business ventures, new banks are more likely to form when business conditions are good. For example, new bank charters bulged to well over 250 per year during the

general economic expansion of the mid-1980s. This high rate of bank start-ups also coincided with the relaxation of unit banking laws in a number of states, laws that had prevented banking companies from operating affiliates in multiple locations. The steady decline in new charters during the late 1980s and early 1990s, which bottomed out at about 50 new banks per year, also had multiple causes. Difficult times in regional banking markets made new bank start-ups unprofitable in many regions (bank failures reached their peak in 1988), and a national recession in the early 1990s reinforced this trend. New bank charters have been on the increase since then, reaching over 100 per year in 1997 and 1998, in large part due to the extended economic expansion of the 1990s.

Conditions in local banking markets also influence bank start-ups. Moore and Skelton (1998) find that there are more de novo banks 1) in markets that are experiencing healthy economic growth, 2) in highly concentrated banking markets in which competition among existing banks is weak, and 3) in markets where small banks are under-represented and, hence, small businesses are not being adequately served. These results imply that new banks will be more likely to start up in local markets where mergers have reduced the number of competing banks, and where the resulting market power has reduced the level of banking services. In such markets, new banks should receive a profitable welcome from customers unhappy with paying high prices for financial services or from businesses whose credit relationships were disrupted when their bank was acquired or failed. Researchers only recently began investigating these phenomena, so there is not yet a consensus on the results. In a study of de novo bank entry in all U.S. markets between 1980 and 1998, Berger, Bonime, Goldberg, and White (1999) find that the probability of de novo entry is higher in local markets that have experienced mergers or acquisitions during the previous three years, particularly mergers and acquisitions involving large banking organizations. In contrast, Seelig and Critchfield (1999) find that local market entry by acquisition deters entry by de novo banks and thrifts.[1] Their results are based on a study of de novo banks and thrifts between 1995 and 1997, a time when banking conditions were exceptional and restrictions on geographic mobility were virtually nonexistent.

Differences in the policies of the legal authorities that grant commercial bank charters can also affect the rate and location of new bank start-ups. A de novo national bank receives its charter from the Office of the Comptroller of the Currency (OCC), while a de novo state bank receives its charter from the banking commission of the home state. The OCC has historically been more liberal in granting charters than most state authorities. Its policy has been that market forces, not the chartering authority, should determine which local markets need and can support new commercial banks. In contrast, many state chartering authorities have historically applied *convenience and needs* tests when considering applications for new bank charters, denying applications if they judge that the convenience and needs of the banking public are already adequately served. Although this federal–state difference in chartering philosophy has diminished over time, DeYoung and Hasan (1998) find that national banks were chartered with greater frequency than state banks during the 1980s and early 1990s, and that the financial performance of de novo national banks initially lagged that of de novo state chartered banks.[2] This suggests that national banks chartered during the 1980s were likely to have had a higher probability of failure than newly chartered state banks operating under similar economic and market conditions.

A concern shared by all chartering authorities is that newly chartered banks start out with enough equity capital to survive through the several years of negative earnings and rapid asset growth that is typical of de novo banks. The dollar amount of start-up financial capital required for approval might be $3 million, $10 million, or even as much as $20 million, depending on the proposed location and business plan of the prospective bank. Larger amounts of start-up capital are generally required for urban banks, for banks locating in vibrant economic markets, and for banks with business strategies that feature fast growth (for example, a new Internet bank).

Once a new bank opens its doors for business, regulatory scrutiny shifts from the applications staff to the examination staff. Bank supervisors pay closer attention to newly chartered banks than to similarly situated established banks, although the difference in treatment varies depending on the new bank's primary regulator. Federal Reserve supervisors will conduct full scope examinations for safety and soundness at a newly chartered bank at six-month intervals (established banks are examined every 12 to 18 months) and will continue to schedule exams at this frequency until the bank receives a strong composite CAMEL rating (that is, a rating of 1 or 2) in two consecutive exams. The Federal Deposit Insurance Corporation requires that all newly chartered state and national banks maintain an 8 percent tier 1 equity capital-to-risk-based assets ratio for their first three years of operation, while the Federal Reserve requires new state chartered Fed member banks to hold this ratio above

9 percent for three years. These temporary extranormal capital requirements for new banks (the tier 1 requirement for established banks to be considered *adequately capitalized* is only 4 percent) are a relatively recent supervisory response to de novo failure experience of the 1980s. Bank supervisors also prohibit new banks from paying out dividends for several years and, in some cases, require new banks to maintain minimum levels of loan loss reserves.

Evolution of new banks

Relatively few research studies have examined how banks grow and evolve in the years immediately after they receive their charters.[3] Brislin and Santomero (1991) show that the financial statements of a new bank can fluctuate rapidly and dramatically during its first year. A handful of studies have examined how the profitability of de novo banks grows over time (for example, Hunter and Srinivasan, 1990, and DeYoung and Hasan, 1998). Another strand of research documents how small business lending becomes less important to de novo banks as they mature (for example, DeYoung, Goldberg, and White, 1999). In this section, I analyze how a broad group of de novo bank characteristics not typically considered in the literature evolve over time, including de novo bank profits, growth rates, capital ratios, sources of income, financing mix, overhead ratios, and loan quality.

Each of the eight panels in figure 2 examines a different financial ratio and compares its average value for a sample of de novo commercial banks to its average value for a sample of established commercial banks. The de novo bank sample includes 4,305 observations of commercial banks that were chartered between 1980 and 1994, were between one and 14 years old when they were observed, and were located in urban banking markets. The established bank sample includes 4,305 observations of commercial banks that were at least 14 years old when they were observed, operated in urban banking markets, and were similar to the de novo banks in terms of asset size. These two samples of banks were originally constructed by DeYoung and Hasan (1998). Box 1 contains additional details about the two bank samples.

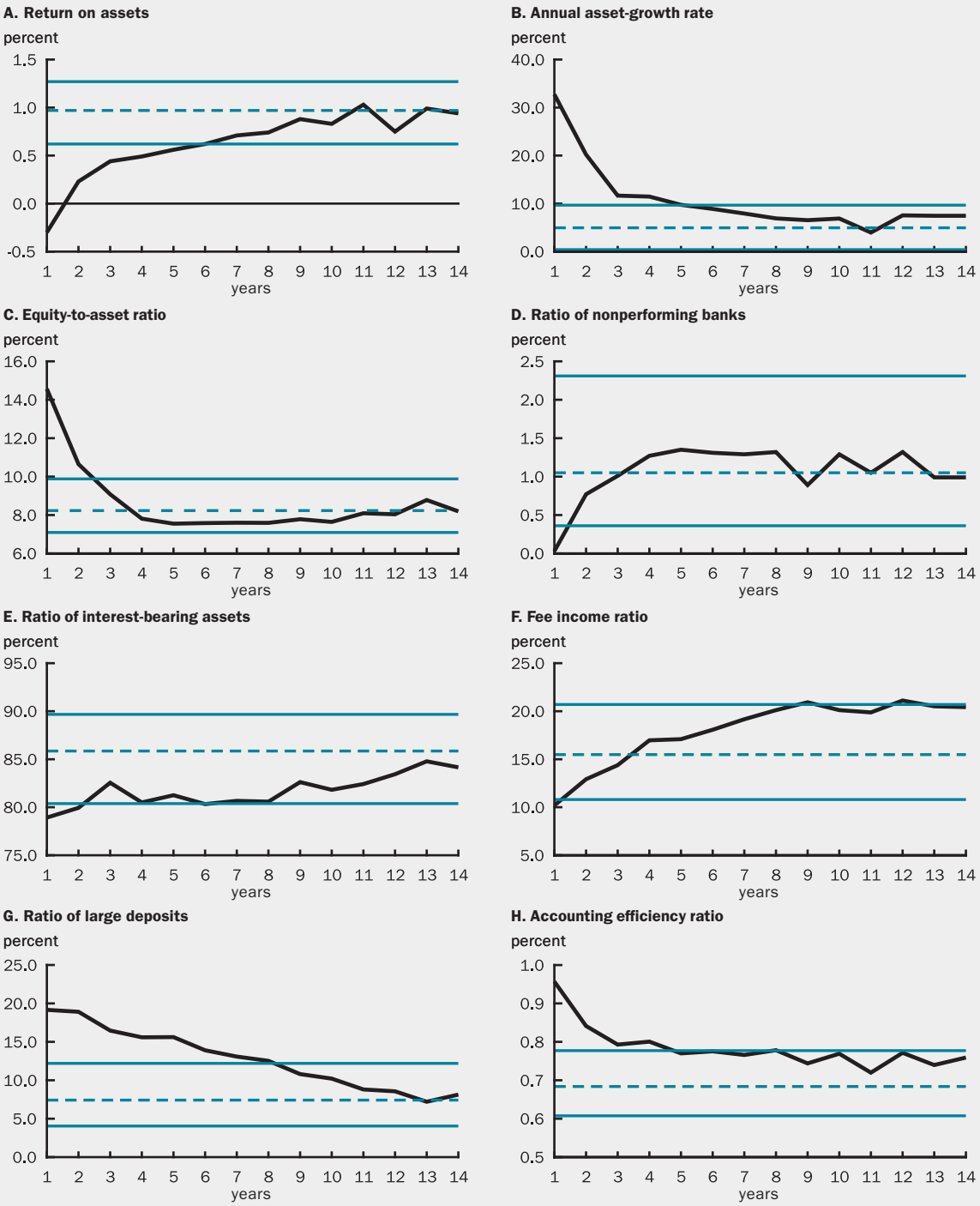
To construct each of the graphs in figure 2, I divided the de novo banks into 14 separate age groups (one-year old banks, two-year old banks, etc.). I then calculated the median average for the financial ratio in question—say, return on assets (ROA)—for each age group. Plotting these 14 average values in chronological order creates a *time path* showing how ROA evolves as the typical de novo bank matures. Finally, I superimposed the value of ROA at the 25th, 50th, and 75th percentiles of the established bank sample as horizontal lines over the de novo bank time path. These horizontal lines serve as *maturity benchmarks* against which to compare the progress of de novo banks over time. The rate at which the de novo time path converges with the maturity benchmarks indicates the speed at which the de novo banks mature.

While each of the individual graphs in figure 2 has a straightforward interpretation when considered in isolation, these eight panels reveal a richer story when they are interpreted in conjunction with each other. For example, by itself the return on assets (ROA) graph (panel A) merely confirms the results of existing studies of de novo bank profitability, that is, that the typical new bank loses money until it is about 18 months old and continues to underperform the average established bank for about a decade. But when the ROA graph is considered together with the asset-growth (panel B) and equity-to-asset (panel C) graphs, a simple theory of de novo failure begins to emerge. De novo banks average an extraordinary 20 percent annual rate of growth during the first three years of their lives. While this fast growth rate is increasing the amount of assets against which new banks need to hold equity capital, the losses suffered during the first and second years of these banks' lives are depleting their equity capital. Despite initially high capital levels, the equity-to-asset ratio of the typical new bank declines very quickly, entering the established bank range after just three years. Thus, panels A, B, and C suggest the probability of failure should increase as new banks pass their third year of life—their capital has declined to established bank levels by year three, but their asset growth and profitability do not converge with those of established banks until at least year ten.

The remaining five panels in figure 2 are consistent with the simple theory of de novo bank failure suggested by the ROA, asset growth, and equity-to-asset panels. For example, newly chartered banks initially have almost no nonperforming loans (panel D). This is because these banks' loan portfolios are composed disproportionately of *unseasoned* loans made recently to borrowers who demonstrated strong financial fundamentals. However, as time passes some of these new borrowers will naturally run into trouble, and the quality of de novo banks' loan portfolios will naturally decline. This happens quite quickly for the typical de novo bank, as its level of nonperforming loans rises slightly above the median level for established banks after three years—just as de novo banks are depleting their excess capital cushions and well before new bank profitability rates have matured.

FIGURE 2

Financial ratio time paths for de novo banks



Notes: The data are described in box 1. The three colored horizontal bands are maturity benchmarks that indicate the twenty-fifth, fiftieth, and seventy-fifth percentiles of the distribution of the ratio in question for the established bank sample. The black line plots the median value of the ratio in question for the banks of various ages in the de novo bank sample. Return on assets is net income divided by total assets. Annual asset-growth rate is the percent increase in total assets over the previous year-end total. Equity-to-asset ratio is the book value of equity divided by total assets. Ratio of nonperforming loans is loans past due 90 or more days plus nonaccruing loans divided by total loans. Ratio of interest-bearing assets is total performing loans plus total securities divided by total assets. Fee income ratio is noninterest income divided by net interest income plus noninterest income. Ratio of large deposits is deposits in accounts greater than $100,000 divided by total deposits. Accounting efficiency ratio is noninterest expense divided by net interest income plus noninterest income.

Source: National Information Center, 1988, 1990, 1992, and 1994, "Report of income and condition," selected banks, December 31.

The slow rate at which de novo bank profitability improves appears to be attributable more to cost factors than to revenue factors. Although the percentage of de novo bank assets invested in interest-bearing assets, such as loans and securities, starts out relatively low and increases only slowly over time (panel E), the typical de novo bank outperforms one-quarter of the established banks in this area after only three years. (De novo bank ROA does not reach the 25th percentile benchmark until year six.) Even more impressive is the speed at which new banks develop the ability to generate fee income (panel F). The typical de novo bank outstrips the average established bank in fee-based revenues after only three years, and outperforms three-quarters of the older banks in this area after about nine years. By virtue of their newness, de novo banks may be less constrained by the inertia of existing customer relationships and existing employee habits and, therefore, may be better able to impose fees on retail customers or to enter into less traditional

BOX 1

Financial ratio time path data

Both the de novo bank sample and the established bank sample were taken from a primary data set used originally in a study by DeYoung and Hasan (1998). For the current study, I added variables from the "Reports of income and condition" (call reports). The primary dataset is an unbalanced panel consisting of 16,282 observations of 5,435 small, urban commercial banks at year-end 1988, 1990, 1992, and 1994. Not all of the banks are present in each of the four years because some banks failed, were acquired, or received their charters during the sample period. There are 2,611 banks present in all four years, 977 banks in three of the four years, 1,005 banks in two years only, and 842 banks in just one year.

Banks had to meet a number of conditions to be included in the primary dataset. First, banks had to have less than $500 million of assets (in 1994 dollars). By definition, newly chartered banks are small, and established banks that are large will not serve as good benchmarks against which to judge the progress of young banks. Large banks have access to production methods, risk strategies, distribution channels, and managerial talent not available to small banks. Second, all banks had to be headquartered in metropolitan statistical areas (MSAs). Demand for banking products, as well as competitive rivalry among banks, can be quite different in rural and urban markets, and may cause young banks to develop differently in these two environments. Third, banks had to be at least 12 months old at the time of observation. For example, a bank that was chartered during 1993, but was observed at year-end 1994, is referred to as a one-year old bank. Brislin and Santomero (1991) find that financial statements are quite volatile during the first year of a bank's operations, which makes performance difficult to measure. Fourth, all banks had to make loans *and* take deposits, eliminating special purpose banks such as credit card banks. Fifth, banks that were 14 years old or less (that is, banks that would be in the de novo sample) were excluded if they held more than $50 million in assets at the end of their first year. This filter prevents established banks that received new charters as part of regulatory reorganizations and established thrift institutions converting to bank charters from being identified as de novo banks.

The resulting de novo sample comprises 4,305 observations of 1,579 different banks 14 years old or younger. Roughly 47 percent of these de novo banks hold federal charters and roughly 21 percent are affiliates in multibank holding companies. The established bank sample was constructed by choosing 4,305 observations of 1,514 different banks, each more than 14 years old, from the primary dataset. Roughly 25 percent of these established banks hold federal charters and roughly 27 percent are affiliates in multibank holding companies. The established banks were chosen to have roughly the same asset-size distribution as the de novo bank sample, as follows: Banks more than 14 years old were grouped into ten asset categories ($0–$50 million, $50–$100 million, ..., $450–$500 million). Established banks were drawn at random from each of these size categories, depending on the number of de novo banks of each asset size. The assets of the resulting established bank sample average $55.97 million with a standard deviation of $49.64 million, compared with the de novo bank sample average of $54.39 million and standard deviation of $48.70 million.

Obviously, there is no bright line that separates de novo banks from established banks. I chose the 14-year old threshold for two reasons. First, it is the maximum age at which previous studies refer to commercial banks as de novo (see Huyser, 1986, and DeYoung and Hasan, 1998). Second, choosing a relatively large number for this threshold ensures that the maturity benchmarks in figure 2 contain only banks that are fully mature.

fee-generating lines of business. In addition, de novo banks tend to start up in markets where business conditions are strong, and selling fee-based financial services may be easier in these markets.

In contrast to their reasonably strong ability to generate revenue, newly chartered banks have a difficult time controlling expenses. De novo banks initially use large deposits twice as intensively as do established banks, and this disparity only slowly disappears (panel G). This suggests that de novo banks tend to finance their fast asset growth by purchasing funds rather than by growing their core deposit base. All else being equal, this is an expensive and potentially risky financing strategy, because large depositors are more sensitive to changes in interest rates than are retail depositors and require higher rates to leave their funds in the bank. The accounting efficiency ratio graph (panel H) indicates that newly chartered banks also have relatively high levels of overhead expenses (for example, branch locations, labor expenses, and computer equipment) and that these fixed factors of production are not used at near full capacity for a number of years. Excess overhead capacity not only depresses bank profitability but, by increasing operating leverage, it makes bank profits more sensitive to fluctuations in bank revenues.

Note that each of the panels in figure 2 exhibits what is known as *survivor bias*, because some de novo banks fail before they are 14 years old. For example, average de novo ROA equals approximately 0.4 percent for the three-year old banks, which is about twice as large as the average ROA of 0.2 percent for the two-year old banks. For the most part, this substantial improvement can be attributed to better performance as young banks grow older and larger. But some amount of this improvement occurs because some of the most unprofitable de novo banks failed between years two and three and dropped out of the sample. Although this second explanation is responsible for only a small amount of the large increase in ROA (as we shall see, very few de novo banks fail after only two years of operation), it is a good illustration of how survivor bias can affect our results. Thus, the most exact way to interpret the ROA time path is as follows: If a newly chartered bank survives to be three years old, one would expect its ROA to be about 0.4 percent. I revisit the issue of survivor bias when I estimate time to failure models in a later section (see Estimating hazard functions section, starting on page 26).

Hypothetical hazard rates

The time paths in figure 2 imply that de novo banks will at first be very unlikely to fail, perhaps even less likely to fail than established banks. Despite the losses typically incurred during their first year of operation, de novo banks initially have very high cushions of equity capital and very low levels of nonperforming loans. But the time paths in figure 2 also imply that de novo banks become dramatically more likely to fail as time passes, and quickly may become more likely to fail than established banks. As de novo banks age, their initially high capital cushions and low nonperforming loan ratios move rapidly toward established bank levels—much more rapidly than their profitability reaches established bank levels.

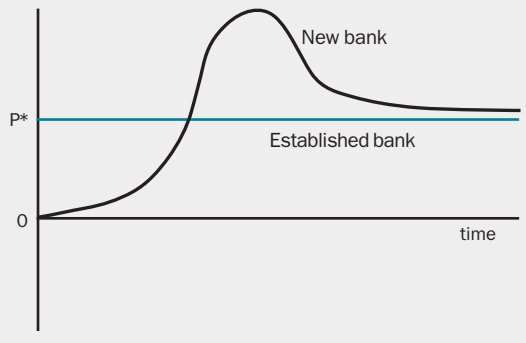
The combined effect of these financial ratio time paths on the timing and probability of de novo bank failure is suggested by the hypothetical *hazard functions* in figure 3. A hazard function tracks changes over time in the *hazard rate*, which is simply the probability that a bank will fail at a particular time, given that it has survived through all of the previous periods leading up to that time.[4] The horizontal line at P* represents the hypothetical hazard rate for established banks, and the curved line plots the hypothetical hazard rate for newly chartered banks. Although this figure is highly stylized, the relative shapes of the two functions are consistent with the combined financial ratio time paths shown in figure 2.

The constant, non-zero hazard rate depicted in figure 3 for established banks is an obvious simplification. Historically, established banks are more prone to failure during recessionary periods, and almost completely unlikely to fail during expansionary periods. This simplification focuses attention on the issue of primary interest here, the failure rate of newly chartered banks *relative to* the failure rate of established banks.

The hypothetical hazard rate for newly chartered banks starts out at zero in figure 3, which makes sense because these banks are so heavily capitalized at the outset. But, as we saw in figure 2, de novo bank capital

FIGURE 3

Hypothetical probabilities of failure



ratios decline to established bank averages after about three years, while de novo bank profits, asset quality, and growth rates do not reach (or return to) established bank levels for around ten years. When these time paths are considered simultaneously, they imply the hypothetical patterns displayed in figure 3. The hypothetical hazard rate for new banks increases at first (for example, between ages one and three) as new banks become increasingly vulnerable to economic fluctuations; it exceeds the established bank hazard rate for a time (for example, after year three); and it eventually declines to converge with established bank levels (for example, around year ten). Regardless of the exact shape and timing of the de novo bank hazard function, it must eventually converge with the established bank hazard function, because by definition new banks that survive eventually turn into established banks.

A rough way to check the relative accuracy of the hypothetical hazard functions drawn in figure 3 is to calculate *Z-score probabilities of failure* for de novo banks and established banks. The Z-scores are constructed as follows:

$$Z = \frac{ROA + equity / assets}{standard\ deviation\ of\ ROA}.$$

The Z-score indicates the number of standard deviations that ROA would have to fall below its average value in order to wipe out 100 percent of the bank's equity capital. For example, if a bank has 5 percent equity capital and, on average, it earns ROA of 1 percent with a standard deviation of 1 percent, then its Z-score would equal 6.00. In this case, the bank's ROA would have to decline by 6 standard deviations below its average (to –5 percent) for its losses to exhaust its capital cushion. Thus, the higher a bank's Z-score, the lower its probability of failure. *Z* will increase (that is, the probability of failure will decrease) with higher levels of average ROA; *Z* will increase with higher levels of equity to assets; and *Z* will increase with lower variability in ROA.[5]

Table 1 displays Z-scores for the established bank sample, for the de novo bank sample, and for several subsamples of de novo banks. All of these calculations employ the data used to construct the graphs in figure 2. For each sample or subsample of banks, *Z* is calculated using the median average of *ROA*, the mean average of *equity/assets*, and the cross-sectional standard deviation of *ROA*. (I use the median *ROA* because the mean *ROA* is skewed downward by banks that incurred large losses.) Because these Z-scores are averages, they represent the likelihood of failure for the typical bank in each sample.

In general, the calculations shown in table 1 suggest that becoming insolvent is a relatively unlikely event for the typical bank in these samples. For example, the lowest Z-score (highest probability of insolvency) is 3.01, or about 3 standard deviations, for the average three- to five-year old bank. Assuming that *Z* is normally distributed, this implies only a 13 in 1,000 (0.13 percent) chance of becoming insolvent. Given the large number of bank failures during the sample period (see figure 1, panel A), the *level* of the failure

TABLE 1

Average Z-scores

	Number of banks	Components of average Z-score: Median ROA	Components of average Z-score: Mean capital-to-assets ratio	Components of average Z-score: Cross-sectional standard deviation of ROA	Average Z-score
De novo banks					
1 to 14 years old	4,305	.0057	.0957	.0230	3.97
Less than 3 years old	667	.0006	.1231	.0206	6.00
3 to 5 years old	1,424	.0050	.0814	.0287	3.01
6 to 10 years old	1,570	.0074	.0762	.0195	4.28
11 to 14 years old	644	.0089	.0795	.0155	5.70
Established banks					
More than 14 years old	4,305	.0097	.0867	.0129	7.48

Notes: Z-scores were calculated using formula described in the text and data described in box 1. The selected commercial banks are also described in box 1.
Source: Author's calculations based on year-end data from the National Information Center, 1988, 1990, 1992, and 1994, "Report of income and condition," selected banks.

probabilities implied by these Z-scores is probably too low.[6] However, these Z-scores are still useful, because they summarize the information in figure 2 into a single number that *ranks* the probability of failure across banks of different ages.

Overall, the analysis suggests that newly chartered banks are more likely to fail than established banks: The average de novo Z-score of 3.97 is considerably smaller than the average established bank Z-score of 7.48. On average, de novo banks and established banks have nearly identical capital-to-asset ratios, so any difference in their implied failure rates must be due to the level and variability of ROA. Indeed, the median de novo bank ROA is only about half as large as the median established bank ROA (.0057 versus .0097), and ROA is nearly twice as variable across the de novo banks than across the established banks (.0230 versus .0129).

Analyzing the Z-scores across de novo banks of different ages provides some support for the shape of the de novo bank hazard function in figure 3. The implied probability of failure is relatively low for banks less than three years old ($Z = 6.00$); is substantially higher for three- to five-year old banks ($Z = 3.01$); and then gradually declines toward established bank levels for banks that survive beyond five years ($Z = 4.28$) and beyond ten years ($Z = 5.70$). Looking at the components of these average Z-scores reveals why the probability of failure changes as new banks mature. The youngest group of de novo banks are the least likely to fail because their earnings are relatively stable (although they average near zero) and their capital cushions are large. The three- to five-year old de novo banks are more likely to fail because, although they have higher average earnings, their capital cushions have been depleted and their earnings are highly variable. Once banks are five to ten years old, increasing earnings, increasing capital, and declining earnings volatility all contribute to a reduced probability of bank failure.

Estimating hazard functions

Next, I test whether the hypothetical hazard functions in figure 3 accurately depict the relative rates at which newly chartered banks and established banks fail. The Z-score analysis discussed above provides some support for these hypothetical hazard functions, but that evidence is crude at best and suffers from survivor bias in the data. In this section, I employ more sophisticated techniques to estimate hazard functions for both newly chartered and established banks. These techniques explicitly account for survivor bias caused by failures and acquisitions during the sample period. In addition, these techniques generate continuous (or nearly continuous) hazard functions that can be plotted against time, making them easy to compare with the shape of the hypothetical hazard functions in figure 3. Finally, one of these techniques tests whether differences in de novo and established bank failure rates are caused by differences in these banks' locational, regulatory, or organizational characteristics.

Data on bank failures

Table 2 displays some summary statistics for a bank failure dataset Federal Reserve Bank of Chicago staff created for the purpose of this study. This dataset contains 56 quarters of information on 2,653 banks from 1985 through 1998, and is constructed from the "Reports of income and condition" (call reports) and from the failures, transformations, and attributes tables in the National Information Center database. The dataset includes 303 newly chartered commercial banks that opened their doors during 1985 and 2,350 established commercial banks that had been in operation for at least 25 years in 1985. The established banks each had less than $25 million in assets (1985 dollars); had equity capital equal to at least 5 percent of their assets; and were located in states in which at least four de novo banks started up in 1985. The dataset tracks each of these 2,653 banks across time and records the quarters in which banks left the dataset because they either failed or were acquired by another bank.

These data cover a period during which there were economic disruptions of sufficient magnitude to cause a statistically meaningful number of bank failures. Commercial bank failures were extremely rare in the U.S. during the 1950s, 1960s, and 1970s, due to generally good economic times, regulatory limits on the risks that banks could take, and legal entry barriers that protected banks from competition. But the combination of banking deregulation and volatile interest rates during the 1970s and 1980s exposed banks to greater risks and more competition. As seen in figure 1, panel A, bank failures accelerated from near zero in 1980 to over 100 failures per year from the mid-1980s through the early 1990s. The catalyst for these bank failures was a series of substantial economic disruptions, including a general recession in the early 1990s and a number of regional recessions in the mid- to late 1980s, the most disruptive of which was due to land price deflations in Texas and other oil-producing states.

For the purposes of this article, I consider a bank to have failed when at least one of the following

TABLE 2

Descriptive statistics for hazard function data sets

	De novo banks	Established banks
Number of banks	303	2,350
Age of banks in 1985 *(years)*	< 1	> 25
Federal charters *(%)*	43.23	84.17
Urban locations *(%)*	78.55	18.81
Multibank holding company *(%)*	32.34	11.62
Southwest states *(%)*, (Texas, Louisiana, Oklahoma)	32.67	10.68
Mean equity/assets	0.368	0.098
Median assets *(current dollars in thousands)*	6,204	14,415
Outcome *(number and % of sample)*		
Failed before 1999	50 (16.5)	185 (7.9)
Acquired before 1999	144 (47.5)	302 (12.9)
Survived to 1999	109 (36.0)	1,863 (79.2)

Notes: The de novo banks began operations during 1985. The established banks were operating in the same states as the de novo banks and were at least 25 years old in 1985. For further details of the data sources and data selection process, see "Estimating hazard functions" section of the text. These data are used to estimate the hazard functions shown in figures 4 and 5.
Sources: Federal Deposit Insurance Corporation, 1985–98, "Report of income and condition," Washington, DC, and National Information Center, 1988, 1990, 1992, and 1994, "Report of income and condition."

occurs: 1) the bank is declared insolvent by its regulator; 2) the bank receives regulatory assistance (for example, a capital injection) without which it would become insolvent; or 3) the bank is acquired soon after its net worth has declined to less than 1 percent of assets. In terms of raw percentages, 16.5 percent of the de novo banks failed before the end of the 14-year sample period. While this is over twice the 7.9 percent failure rate for the established banks in the sample, it is well below the reported failure rates for new (nonbank) business ventures. (See box 2 for a short discussion of new bank failures versus new business failures.)

Both the sample de novo banks and the sample established banks were more likely to be acquired than to fail during the sample period. The new banks were more likely to hold state charters; to be located in urban areas; to be located in the Southwest (primarily Texas, but also Louisiana and Oklahoma); and to be affiliates in multibank holding companies. Some of the hazard functions I estimate below include tests of whether these locational, organizational, and regulatory characteristics affect the probability of bank failure.

Nonparametric hazard functions

I use the bank failure data, summarized in table 2, to estimate separate hazard functions for newly chartered banks and established banks, and then compare these estimated hazard functions with the hypothetical hazard functions in figure 3. I employ two different hazard function techniques to produce these estimates—a *nonparametric*, or *actuarial*, approach, and a *parametric*, or *duration model*, approach.

An actuarial hazard function is simply a series of actuarial hazard rates strung together in chronological order. Calculating the actuarial hazard rates is straightforward and intuitive. For example, to calculate the 1990 hazard rate for a set of banks that were chartered in 1985, one simply divides the number of these banks that failed during 1990 by the number of the banks that still existed at the beginning of 1990. Thus, the hazard rate tells us the probability of failure in 1990 conditional on having survived for five years. The following, more exact, formula can be used to calculate the actuarial hazard rate for any time period, T:

$$hazard(T) = \frac{no.\ of\ bank\ failures\ during\ T}{no.\ of\ banks\ surviving\ at\ start\ of\ T} \approx \frac{f(T)}{n(t=0) - \sum_{t=0}^{T-1}\left(f(t)+m(t)\right) - \frac{1}{2}m(T)},$$

where $n(t = 0)$ is the number of banks present at the beginning of the analysis; $f(t)$ represents the number of these banks that failed during time period t; $m(t)$ represents the number of these banks that were acquired in mergers during time period t, and T indicates the current time period. Note the subtle adjustment to the denominator in the second line of this formula: The denominator is reduced by one-half the number of banks that were acquired during the current time period. These banks clearly did not survive until the end of time period T, and subtracting some portion of these banks from the denominator acknowledges the possibility that they might have failed during time T

BOX 2

New bank failures and new business failures

During the 14 years covered by the bank failure dataset (see table 2), 16.5 percent of the newly chartered banks failed, compared with only 7.9 percent of the established banks of comparable size and location. To put these new bank failure rates into perspective, note that a 16.5 percent failure rate over 14 years is substantially lower than the failure rates typically reported for business start-ups in general. Raw data reported by the U.S. Small Business Administration (1992) suggest that at least 60 percent of new business ventures with less than 500 employees that started in 1977–78 failed within six years. Kirchhoff (1994, pp. 153–169) argues convincingly that these raw data overstate the new business failure rate because, among other things, the data in many instances define firms that changed owners or voluntarily shut down as having failed. After adjusting for these and other factors, Kirchhoff concludes that, in a best case scenario, 18 percent of new business ventures fail within eight years of start-up—about the same rate of failure as the de novo banks but in half the number of years. Furthermore, the 16.5 percent failure rate for new banks occurred during the worst period of bank failures since the Great Depression.

It should not be surprising that new banks have a better rate of survival than other new businesses. Both federal and state bank regulators deny charters to applicants with questionable financial credentials, restrict business activities, require high amounts of capital, apply regular scrutiny via on-site exams, and have the power to revoke bank charters. Banking start-ups face more severe entry barriers and ongoing scrutiny than new businesses in most other industries, and this *selection bias* naturally leads to a higher survival for new banks.

had they not been acquired. Although weighting these banks by one-half is a crude and *ad hoc* adjustment, it is important to make some kind of adjustment because, as shown in table 2, acquired banks greatly outnumbered failed banks between 1985 and 1998.

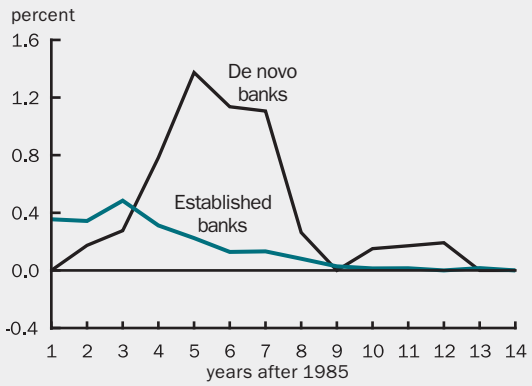
I use the above formula to calculate 14 separate hazard rates (one rate for each of the 14 years from 1985 through 1998) for the 303 newly chartered banks. I repeat this exercise for the 2,350 established banks. Plotting the resulting hazard rates in chronological order generates two nonparametric hazard functions, which are displayed in figure 4.

In general, the nonparametric hazard functions in figure 4 resemble the hypothetical hazard functions posited in figure 3. The hazard rate for newly chartered banks is initially zero, and it remains below the established bank hazard rate for several years. As discussed above, this is most likely because the typical new bank holds a healthy equity cushion at the outset. After year three, the new bank hazard rate exceeds the established bank hazard rate, and it remains substantially higher than the established bank hazard rate until year eight. The hazard rate for newly chartered banks peaks in years five, six, and seven at about 1.2 percent—that is, if a newly chartered bank reaches the beginning of any of these years without failing or being acquired, it has about a 1.2 percent chance of failing before the year is out. At this point, the typical new bank's capital ratio has declined to established bank levels, but its profitability has not yet attained the level or degree of stability found at established banks. After year eight, the new bank hazard rate approaches the established bank hazard rate from above, suggesting that the maturation of new banks is well under way at this point.

The results displayed in figure 4 are consistent with the simple life-cycle theory of de novo bank failure. The nonparametric techniques used to generate figure 4 paint a good general picture of the rate at which new banks fail *relative to* established banks. But these nonparametric techniques do not control for the survivor bias in the data and, as a result, they understate the hazard rate at any given point in time.

FIGURE 4

Nonparametric hazard rates



Furthermore, these techniques are not useful for testing how much, if any, of the difference between the new bank and established bank failure rates is caused by the economic, regulatory, and organizational conditions under which newly chartered banks operate.
In the final step in this analysis, I use econometric *duration analysis* to estimate hazard functions. These parametric methods account for survivor bias and control for environmental conditions that can affect the probability of failure.

Parametric hazard functions

Duration analysis is a statistical regression approach. The dependent variable in these regressions is t, the length of time that passes between a new bank's start-up date and its subsequent failure. For established banks, t is the length of time between its first observation in the dataset (in this case, the first quarter of 1985) and its subsequent failure. The period measured by t is often referred to as a bank's *duration*. Because the banks in this dataset are observed quarterly, duration will range from $t = 1$ for banks that fail during the quarter in which they begin operations, to $t = 56$ for banks that fail in the fourth quarter of 1998.

The simplest duration approach includes no explanatory variables. The analyst starts by selecting a probability distribution formula that has a shape that is roughly similar to the actual distribution of the duration variable t, and uses maximum likelihood techniques to estimate parameter values that shape that probability distribution formula more exactly to the actual duration data. Here, I use a *log-logistic* distribution formula, because this is capable of producing hazard functions that have shapes similar to the hazard functions in figures 3 and 4. (Details of these duration model procedures can be found in the appendix to this article or in Greene, 1997). Once the parameters of the distribution formula have been estimated, they can be used to construct hazard functions as follows:

$$hazard(T) = \frac{f(T)}{1 - F(T)},$$

where $f(T)$ is the probability that a bank fails at time T (that is, the log-logistic probability density) and $F(T)$ is the probability that a bank fails before time T (that is, the log-logistic cumulative probability distribution). The denominator, $1 - F(T)$, is the log-logistic *survival function*, which is the probability that a bank neither fails nor is acquired before time T. This parametric hazard function has the same general interpretation as the nonparametric hazard function calculated in the previous section—they are both estimates of the probability that a bank will fail at time T given that it has survived until time T. One difference is that the hazard function generated by this parametric approach will be a smooth and continuous function of time similar to the hypothetical hazard functions in figure 3, as opposed to the segmented nonparametric hazard function in figure 4.

The duration models I estimate here control for survivor problems in the data. Recall that many of the sample banks either survived beyond the end of the sample period or were acquired during the sample period. These banks are known as *censored observations*. We cannot assign a duration value t to these banks because we cannot observe their ultimate fate (failure or survival). Furthermore, history suggests that very few of these banks will eventually fail, so including them in hazard rate calculations creates a downward bias by inflating the survival function $1 - F(t)$. Duration models can adjust for this problem by estimating the probability that censored banks will eventually fail, and then weighting the censored observations by this probability before estimating the parameters of the hazard function. (See the appendix for more details.)

The sample banks differ in terms of their geographic location, their organizational form, and their primary regulator. These characteristics could make a bank more or less likely to fail, or given that a bank does fail, these characteristics could influence how quickly it fails. For example, banks located in depressed economic regions will be more likely to fail and, absent regulatory intervention, will fail more quickly than banks located in economically healthy markets. Duration models can include a vector of independent variables, typically known as *covariates*, measuring the characteristics that vary across banks but remain constant for each bank over the sample period. I use a split population approach which estimates two regression coefficients for each of the covariates in the duration model. The first coefficient measures the covariate's impact on the probability that a bank will survive—a negative coefficient indicates that the covariate is associated with a lower probability of survival (higher probability of failure). The second coefficient measures the covariate's impact on a bank's duration—given that a bank will eventually fail, a negative coefficient indicates that the covariate is associated with a shorter duration (a faster failure).

The duration models I estimate include four covariates, each of which is expressed as a (0, 1) dummy variable. $OCC = 1$ if the bank holds a federal charter (as opposed to a state charter). The OCC has traditionally practiced a more lenient chartering policy than most state chartering authorities, relying on market

forces rather than administrative rules to determine the number of banks a market could support.[7] A negative coefficient on *OCC* would suggest that this policy caused new national banks to fail more often and/or more quickly, on average, than new state-chartered banks. *INDEPENDENT* = 1 if the bank is either a free-standing business or a one-bank holding company (as opposed to being an affiliate of a multibank holding company) throughout the sample period. A negative coefficient on *INDEPENDENT* would suggest that banks not having access to the financial strength and managerial expertise of a multibank holding company tend to fail more often and/or more quickly. *MSA* = 1 if the bank is located in an urban area. Banks in urban areas face greater competition than rural banks, but also may have greater opportunities for diversification. A negative coefficient on *MSA* would suggest that, on balance, these conditions cause banks in urban areas to fail more often and/or more quickly than rural banks. *SW* = 1 if the bank is located in the southwestern states of Louisiana, Texas, or Oklahoma, which experienced large numbers of bank failures during the mid- to late 1980s due to disruptions in energy-related industries. One would expect the coefficients on *SW* to be negative, reflecting lower survival probabilities and shorter duration times for banks in this region.

I add these four covariates to the duration model merely to illustrate how conditions and events external to the bank can affect its probability of failure and its time to failure. These four variables are not meant to be an exhaustive list of such conditions. Similarly, the duration model I estimate here is by no means definitive of the duration model techniques available to researchers. Other duration approaches do exist, including those that allow for *time-varying covariates* (for example, changes in economic, regulatory, or competitive conditions during each bank's duration). However, the multiple approaches I employ (including the Z-score and actuarial hazard function analysis conducted above) serve the purpose of this study, which is to test the simple life-cycle theory of de novo bank failure summarized in figure 3.

Table 3 displays the results of the duration models estimated separately for newly chartered banks and established banks. The estimated probability that the average bank will eventually fail is 19.65 percent for de novo banks and 8.93 percent for established banks. Note that these estimated failure probabilities are somewhat higher than the raw failure percentages shown at the bottom of table 2. In each case, the estimated probability is higher than the raw percentage because of the possibility that some of the censored observations will eventually fail.

Although established banks are less likely to fail, those that do fail have relatively short durations. Of the established banks that are expected to eventually fail, half of them will fail within an estimated 9.8 quarters (about 2.5 years) after the beginning of the sample period. Consistent with the life-cycle theory, newly chartered banks fail more slowly than established banks. It takes an estimated 21.1 quarters (about 5.25 years) for half of the de novo banks that are expected to fail to do so.

These differences in average duration can be seen clearly in figure 5, which charts the estimated hazard rates from the de novo and established bank duration models. Each of these functions is plotted based on the estimated coefficients shown in table 3 and the average values of the covariates for each sample. In general, these two estimated hazard functions resemble the shapes displayed above in figures 3 and 4. Thus, after controlling for censored data and a variety of environmental conditions, the failure patterns of newly chartered banks still differ substantially from the failure patterns of established banks.

The estimated probability of failure for established banks starts out above zero; peaks at about 8 percent for banks that survive for two years; and then slowly declines as the bank failure wave dissipates (see figure 1). In contrast, the estimated probability of failure for de novo banks starts out at zero and remains lower than the established bank hazard rate for three years; increases rapidly and peaks at nearly 14 percent for banks the survive for seven years; and then declines relatively quickly and begins to approach the established bank hazard rate. Note that both of these hazard functions peak much higher on the vertical scale than did the actuarial hazard functions plotted in figure 4. Thus, by not controlling for censored observations and the overall low probability of eventual failure, the actuarial model substantially understated the hazard rates. Also, note that the hazard rates in figure 5 are in decline but are still positive at year 14, which reflects the non-zero probability of failure for the censored observations.

As expected, being located in one of the southwestern states reduces the probability of survival (or increases the probability of failure) for both de novo and established banks. Failing de novo banks also failed more quickly in this region, but failing established banks had longer than average durations. The latter result may indicate that regulators allowed troubled banks with longstanding business relationships (and, hence, more franchise value) more time to recover before stepping in to resolve them.[8] Being located in a metropolitan statistical area reduced the probability

TABLE 3

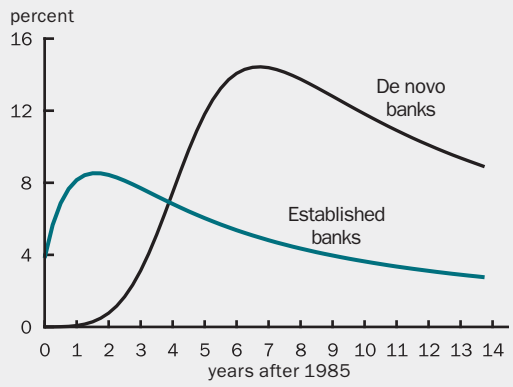
Selected results from parametric duration models

	De novo banks	Established banks
Number of banks in sample	303	2,350
Average predicted failure probability	19.65%	8.93%
Predicted time for 50 percent of banks to fail	21.1 quarters	9.8 quarters
Probability of survival parameter estimates		
Constant	1.2930** (0.5144)	1.2475*** (0.2145)
OCC (= 1 if national bank)	0.0197 (0.2242)	–0.1296 (0.1117)
SW (= 1 if in Louisiana, Oklahoma, or Texas)	–0.3928* (0.2219)	–0.7419*** (0.0992)
MSA (= 1 if urban bank)	–0.5266* (0.3205)	0.2530** (0.1163)
INDEPENDENT (= 1 if independent or sole bank in a one-bank holding company)	–0.5265** (0.2386)	#
Survival time parameter estimates		
Constant	3.3045*** (0.6124)	2.3369*** (0.4489)
OCC	0.0075 (0.1376)	0.1354 (0.2339)
SW	–0.3243** (0.1386)	0.4753** (0.1971)
MSA	–0.2204 (0.5318)	0.5391** (0.2699)
INDEPENDENT	–0.1195 (0.1547)	#

*, **, and *** indicate significance at the 10 percent, 5 percent, and 1 percent levels, respectively.
Notes: Both models are estimated using the data sets described in table 2. Standard errors are in parentheses. # indicates that it was necessary to exclude the variable INDEPENDENT to make the established bank model converge. Further details on these models can be found in the appendix to this article.

FIGURE 5

Parametric hazard functions



of survival for de novo banks, but increased both the probability of survival for established banks and the survival time for established banks likely to eventually fail. Recall that intense competitive rivalry can cause banks to fail in urban markets, and that the lack of diversification opportunities can cause banks to fail in rural markets. The results suggest that these two phenomena affect de novo banks and established banks differently—on balance, de novo banks may be more sensitive to competition than to diversification risk, while small established banks may be more affected by a lack of diversification than by competitive rivalry. Being an independent bank or banking organization also reduces the probability of survival for de novo banks, which suggests that having access to the resources of multibank holding companies helps

new banks survive. (I excluded this covariate from the established bank model because its presence prevented the model from converging.) The identity of a bank's primary regulator (OCC or state) is not a significant determinant of the probability of survival or the survival time for either set of banks.

Conclusion

Like all new business ventures, banks start with a business plan but no guarantee of success. So, despite the regulatory safeguards of on-site examinations, capital requirements, and other risk controls, we should not be surprised to find that new banks are more likely to fail than established banks. This article offers a simple framework that explains not only *why* but also *when* new banks are likely to fail.

My results suggest that the primary determinant of new bank failure is *how new* the bank is. Ironically, de novo banks are relatively unlikely to fail during their first few years of operation when they are earning negative profits. They are relatively more likely to fail during the years of positive profits that follow. Brand new, but unprofitable, banks are typically protected from failure by large initial capital cushions. However, equity cushions at de novo banks typically decline to established bank levels *several years before* their earnings become stable enough to justify these relatively low levels of capital.

What are the implications of this result for capital regulation at newly chartered banks? If ensuring a high rate of survival for de novo banks is a regulatory objective, then this result offers support for requiring high levels of start-up capital for new banks, and for holding young banks to higher capital requirements. Higher levels of required capital will make newly chartered banks less vulnerable to failure. Under such policies, de novo entrants might be a more credible long-run deterrent to market power in consolidating local markets. Indeed, in the wake of the wave of de novo bank failures during the 1980s, federal bank supervisory agencies did impose higher capital requirements on newly chartered banks.

On the other hand, promoting the safety and soundness of the banking system does not require that regulators prevent all bank failures, much less all failures of new banks. At some point, attempting to improve the survival rate of de novo banks by increasing the amount of capital necessary for investors to secure the charter will act as an entry barrier. Similarly, increasing the required capital ratios for young banks with charters already in hand will, at some point, depress investors' expected rates of return and discourage investment in new banks. Higher capital requirements for young banks could also slow the rate at which they can grow their balance sheets, hampering the beneficial impact of new banks in markets where existing banks (perhaps with market power) are not adequately serving the banking public.

What are the implications of this study for the bank chartering decision? During the period covered by this study, some state chartering authorities would approve or deny a charter application only after considering whether a local market "needed" an additional bank, based on the number of banks already serving the market and the expected rate of local economic growth. These restrictive chartering policies sought to reduce bank failure rates, and the financial disruptions that accompany them, by limiting competition in local banking markets. In contrast, the federal chartering authority practiced a liberal entry policy that explicitly ignored these "convenience and needs" issues, stressing instead the potential procompetitive benefits of de novo entry. My results indicate that the de novo national banks chartered in 1985 were no more likely to fail, or to fail quickly, than the de novo state banks chartered in that same year. This suggests that the benefits of a liberal chartering policy can be achieved without substantial increases in de novo bank failure rates. Additional research might confirm whether these findings, which are based on data from just 303 new banks chartered in a single year, also hold for banks chartered in other years and/or under different economic and regulatory circumstances.

APPENDIX

Split population duration models

The parametric hazard functions described in the text begin with the assumption that a population of N banks will fail over time period $(0,t)$ according to some probability distribution:

$$F(t) = \int_0^t f(t)dt,$$

where t represents time and $f(t)$ is the probability density function associated with $F(t)$. The hazard function $h(t)$ can then be written as a function of $F(t)$ and $f(t)$ as follows:

$$h(T) = \frac{f(T)}{1-F(T)} = \frac{f(T)}{S(T)},$$

where $S(t) = 1 - F(t)$ is the survival function and $0<T<t$. Thus, $h(T)$ gives the probability (that is, the hazard rate) that a bank will fail at T conditional on surviving until T.

The general shape of the estimated hazard function will depend on the underlying probability distribution chosen to fit the data. I use the log-logistic distribution because it is capable of producing the hazard function shapes hypothesized in figure 3. The log-logistic distribution imposes the following functional forms on the hazard and survival functions:

$$h(t) = \frac{\lambda p(\lambda p)^{p-1}}{1+(\lambda t)^p}$$

$$S(t) = \frac{1}{1+(\lambda t)^p},$$

where the parameters p and λ give the hazard function its exact shape. The parameter p captures *duration dependence* or whether the hazard rate increases or decreases across time. The parameter λ captures the portion of the hazard rate that is time-invariant. This parameter, which can take on different values for different banks, is expressed as follows:

$$\lambda_i = e^{-\beta' X_i},$$

where the bank index i ranges from 1 to N, and X_i is a vector of bank-specific covariates that do not vary over time. Table 3 reports the estimated values of β under the heading "survival time parameter estimates." I use these β estimates to evaluate the above expression at the means of the covariates, which results in $\lambda = 0.0474$ for the average de novo bank and 0.1019 for the average established bank. The parameter p is a constant that does not vary across banks; it equals 5.0175 for the de novo bank model and 1.6333 for the established bank model.

All of the parameters of the duration models in this study were estimated using maximum likelihood techniques. The standard estimation procedure for duration models starts with the following likelihood function:

$$L = \prod_{i=1}^{N}\left[f(t_i|p,\beta)\right]^{Q_i}\left[S(t_i|p,\beta)\right]^{1-Q_i},$$

where $Q_i = 1$ if bank i failed during the sample period (an uncensored observation) and $Q_i = 0$ if bank i survived or was acquired during the sample period (a censored observation). Substituting $h(t)/S(t)$ for $f(t)$ and performing a log transformation produces the log-likelihood function to be maximized:

$$\ln L = \sum_{i=1}^{UC}\ln h(t_i|p,\beta) + \sum_{i=1}^{N}\ln S(t_i|p,\beta),$$

where $i \le UC$ are the uncensored observations. Once I have estimated the parameters p and β, I can calculate the median time to failure by setting $S(t) = 0.50$ and solving for t.

This standard approach is based on the assumption that all of the censored banks will eventually fail (or would have eventually failed had they not been acquired). This assumption is inappropriate for the data used here, however, because over 80 percent of the de novo banks were censored observations, and over 90 percent of the established banks were censored observations. Given the nature of the data, I use a more general framework that avoids making this assumption. In the *split population duration model*, an additional estimable parameter δ, the probability that a bank eventually fails, enters the likelihood function as follows:

$$L = \prod_{i=1}^{N}\left[\delta f(t_i|p,\beta)\right]^{Q_i}\left[(1-\delta) + \delta S(t_i|p,\beta)\right]^{1-Q_i}.$$

Both Cole and Gunther (1995) and Hunter, Verbrugge, and Whidbee (1996) estimate split population models of financial institution failure. Note that this formulation collapses to the standard framework when $\delta = 1$. But when $\delta<1$, the functions $S(t)$ and $f(t)$ become *conditional on* the bank eventually failing. Thus, the estimated hazard function $h(t) = f(t)/S(t)$

will not be unduly influenced by censored observations of banks that have little chance of ever failing. The parameter δ can vary across banks as a function of a bank's covariate values:

$$\delta_i = \frac{1}{1 + e^{\alpha' X_i}}.$$

Table 3 reports the estimated values of α under the heading is "probability of survival parameter estimates." I use these α estimates to evaluate the above expression at the means of the covariates, which results in δ = 0.1965 for the average de novo bank and 0.0893 for the average established bank.

The hazard functions plotted in figure 5 show the probability that the average bank will fail at time *t*, given that the bank has not yet failed *but will eventually fail*. Thus, the shapes of the plotted hazard functions are based on the estimated values of λ and *p*, but not on the estimated values of δ.

NOTES

[1]Note that such ambiguity is largely absent from studies that examine the determinants of local market entry by already established banks. For a recent example, see Amel and Liang (1997). In general, these studies tend to find that established banks are more likely to enter highly profitable local banking markets, but less likely to enter highly concentrated local banking markets. Of course, the causes and consequences of a new bank start-up may be quite different from the causes and consequences of market entry by an already established bank.

[2]Seelig and Critchfield (1999) find that, on average, the state chartering authorities remain relatively more likely than the OCC to consider the ability of local banking markets to support an additional bank when evaluating a charter application. The authors show that income per capita per branch in the local banking market was a substantially stronger predictor of de novo state bank entry than of de novo national bank entry between 1995 and 1997.

[3]See DeYoung and Hasan (1998) for a more complete review of this literature.

[4]Examples of studies that have used hazard rates to analyze financial institution failure include Whalen (1991); Wheelock and Wilson (1995); Cole and Gunther (1995); Helwege (1996); and Hunter, Verbrugge, and Whidbee (1996).

[5]The Z-score is a measure of the probability that a firm's losses (negative profits) will exceed its equity capital. See Brewer (1989) for a discussion of the Z-score and its use in banking research.

[6]There are a number of possible reasons for this. In general, Z-score analysis performs best when used to represent the likelihood that an individual firm will become insolvent and, as such, Z-scores are typically constructed based on the known distribution (the mean and standard deviation) of ROA for an individual firm. In contrast, these average Z-scores are constructed for groups of banks, and rely on the cross-sectional distribution (the median and standard deviation) of ROA for each group of banks. As mentioned in the text, the distribution of ROA is not normally distributed, but rather is relatively skewed. Hence, it would be inappropriate to use the absolute levels of these average Z-scores to draw statistical inferences about the probability of bank failure.

[7]See Hunter and Srinivasan (1990) and DeYoung and Hasan (1998) for discussions that compare historical federal and state chartering policies.

[8]There are many potential reasons for the high bank failure rates in Texas, and for the relatively shorter durations for de novo banks in Texas, during the 1980s and 1990s. These reasons include unexpected economic shocks, unit banking restrictions that limited geographic diversification, a relatively undiversified regional economy, and regulatory failure.

REFERENCES

Amel, Dean F., and J. Nellie Liang, 1997, "Determinants of entry and profits in local banking markets," *Review of Industrial Organization*, Vol. 12, No. 1, pp. 59–78.

Berger, Allen N., Seth D. Bonime, Lawrence G. Goldberg, and Lawrence J. White, 1999, "The dynamics of market entry: The effects of mergers and acquisitions on de novo entry and customer service in banking," *Proceedings from a Conference on Bank Structure and Regulation*, Federal Reserve Bank of Chicago, forthcoming.

Brewer III, Elijah, 1989, "Relationships between bank holding company risk and nonbank activity," *Journal of Economics and Business*, Vol. 41, November, pp. 337–353.

Brislin, Patricia, and Anthony Santomero, 1991, "De novo banking in the Third District," *Business Review*, Federal Reserve Bank of Philadelphia, January, pp. 3–12.

Cole, Rebel A., and Jeffery W. Gunther, 1995, "Separating the likelihood and timing of bank failure," *Journal of Banking and Finance*, Vol. 19, No. 6, September, pp. 1073–1089.

DeYoung, Robert, Lawrence G. Goldberg, and Lawrence J. White, 1999, "Youth, adolescence, and maturity of banks: Credit availability to small business in an era of banking consolidation," *Journal of Banking and Finance*, Vol. 23, No. 2-4, February, pp. 463–492.

DeYoung, Robert, and Iftekhar Hasan, 1998, "The performance of de novo commercial banks: A profit efficiency approach," *Journal of Banking and Finance*, Vol. 22, May, pp. 565–587.

Greene, William H., 1997, *Econometric Analysis*, Upper Saddle River, NJ: Prentice Hall.

Helwege, Jean, 1996, "Determinants of savings and loan failure," *Journal of Financial Services Research*, Vol. 10, No. 4, pp. 373–392.

Hunter, William C. and Aruna Srinivasan, 1990, "Determinants of de novo bank performance," *Economic Review*, Federal Reserve Bank of Atlanta, March, pp. 14–25.

Hunter, William C., James A. Verbrugge, and David A. Whidbee, 1996, "Risk taking and failure in de novo savings and loans in the 1980s," *Journal of Financial Services Research*, Vol. 10, No. 3, pp. 235–272.

Huyser, Daniel, 1986, "De novo bank performance in the Seventh District states," *Banking Studies*, Federal Reserve Bank of Kansas City, pp. 13–22.

Kirchhoff, Bruce A., 1994, *Entrepreneurship and Dynamic Capitalism*, Westport, CT and London: Greenwood, Praeger.

Moore, Robert R., and Edward C. Skelton, 1998, "New banks: Why enter when others exit?," *Financial Industry Issues*, Federal Reserve Bank of Dallas, First Quarter, pp. 1–7.

Seelig, Stephen, and Timothy Critchfield, 1999, "Determinants of de novo entry in banking," Federal Deposit Insurance Corporation, working paper, No. 99-1, January.

U.S. Small Business Administration, 1992, *The State of Small Business*, Washington, DC: U.S. Government Printing Office.

Whalen, Gary, 1991, "A proportional hazards model of bank failure: An examination of its usefulness as an early warning tool," *Economic Review*, Federal Reserve Bank of Cleveland, Vol. 27, No. 1, First Quarter, pp. 20–31.

Wheelock, David C., and Paul W. Wilson, 1995, "Explaining bank failures: Deposit insurance, regulation, and efficiency," *Review of Economics and Statistics*, November, pp. 689–700.

New facts in finance

John H. Cochrane

Introduction and summary

The last 15 years have seen a revolution in the way financial economists understand the investment world. We once thought that stock and bond returns were essentially unpredictable. Now we recognize that stock and bond returns have a substantial predictable component at long horizons. We once thought that the capital asset pricing model (CAPM) provided a good description of why average returns on some stocks, portfolios, funds, or strategies were higher than others. Now we recognize that the average returns of many investment opportunities cannot be explained by the CAPM, and "multifactor models" are used in its place. We once thought that long-term interest rates reflected expectations of future short-term rates and that interest rate differentials across countries reflected expectations of exchange rate depreciation. Now, we see time-varying risk premiums in bond and foreign exchange markets as well as in stock markets. We once thought that mutual fund average returns were well explained by the CAPM. Now, we see that funds can earn average returns not explained by the CAPM, that is, unrelated to market risks, by following a variety of investment "styles."

In this article, I survey these new facts, and I show how they are variations on a common theme. Each case uses price variables to infer market expectations of future returns; each case notices that an offsetting adjustment (to dividends, interest rates, or exchange rates) seems to be absent or sluggish. Each case suggests that financial markets offer rewards in the form of average returns for holding risks related to recessions and financial distress, in addition to the risks represented by overall market movements. In a companion article in this issue, "Portfolio advice for a multifactor world," I survey and interpret recent advances in portfolio theory that address the question, What should an investor do about all these new facts?

First, a slightly more detailed overview of the facts then and now. Until the mid-1980s, financial economists' view of the investment world was based on three bedrocks:

1. The CAPM is a good measure of risk and thus a good explanation of the fact that some assets (stocks, portfolios, strategies, or mutual funds) earn higher average returns than others. The CAPM states that assets can only earn a high average return if they have a high "beta," which measures the tendency of the individual asset to move up or down with the market as a whole. Beta drives average returns because beta measures how much adding a *bit* of the asset to a diversified portfolio increases the volatility of the *portfolio*. Investors care about portfolio returns, not about the behavior of specific assets.

2. Returns are unpredictable, like a coin flip. This is the *random walk* theory of stock prices. Though there are bull and bear markets; long sequences of good and bad *past* returns; the expected *future* return is always about the same. *Technical analysis* that tries to divine future returns from patterns of past returns and prices is nearly useless. Any apparent predictability is either a statistical artifact which will quickly vanish out of sample or cannot be exploited after transaction costs.

Bond returns are not predictable. This is the *expectations model* of the term structure. If long-term bond yields are higher than short-term yields—if the yield curve is upward sloping—this does not mean that you expect a higher return by holding long-term bonds rather than short-term bonds. Rather, it means

John H. Cochrane is the Sigmund E. Edelstone Professor of Finance in the Graduate School of Business at the University of Chicago, a consultant to the Federal Reserve Bank of Chicago, and a research associate at the National Bureau of Economic Research (NBER). *The author thanks Andrea Eisfeldt for research assistance and David Marshall, John Campbell, and Robert Shiller for comments. The author's research is supported by the Graduate School of Business and by a grant from the National Science Foundation, administered by the NBER.*

that short-term interest rates are expected to rise in the future. Over one year, the rise in interest rates will limit the capital gain on long-term bonds, so they earn the same as the short-term bonds over the year. Over many years, the rise in short rates improves the rate of return from rolling over short-term bonds to equal that of holding the long-term bond. Thus, you expect to earn about the same amount on short-term or long-term bonds at any horizon.

Foreign exchange bets are not predictable. If a country has higher interest rates than are available in the U.S. for bonds of a similar risk class, its exchange rate is expected to depreciate. Then, after you convert your investment back to dollars, you expect to make the same amount of money holding foreign or domestic bonds.

In addition, stock market volatility does not change much through time. Not only are returns close to unpredictable, they are nearly identically distributed as well. Each day, the stock market return is like the result of flipping the same coin, over and over again.

3. Professional managers do not reliably outperform simple indexes and passive portfolios once one corrects for risk (beta). While some do better than the market in any given year, some do worse, and the outcomes look very much like luck. Funds that do well in one year are not more likely to do better than average the next year. The average actively managed fund performs about 1 percent *worse* than the market index. The more actively a fund trades, the lower the returns to investors.

Together, these views reflect a guiding principle that asset markets are, to a good approximation, *informationally efficient* (Fama, 1970, 1991). Market prices already contain most information about fundamental value and, because the business of discovering information about the value of traded assets is extremely competitive, there are no easy quick profits to be made, just as there are not in any other well-established and competitive industry. The only way to earn large returns is by taking on additional risk.

These views are not ideological or doctrinaire beliefs. Rather, they summarize the findings of a quarter century of careful empirical work. However, every one of them has now been extensively revised by a new generation of empirical research. The new findings need not overturn the cherished view that markets are reasonably competitive and, therefore, reasonably efficient. However, they do substantially enlarge our view of what activities provide rewards for holding risks, and they challenge our understanding of those risk premiums.

Now, we know that:

1. There are assets whose average returns can not be explained by their beta. Multifactor extensions of the CAPM dominate the description, performance attribution, and explanation of average returns. Multifactor models associate high average returns with a tendency to move with other risk factors in addition to movements in the market as a whole. (See box 1.)

2. Returns are predictable. In particular: Variables including the dividend/price (d/p) ratio and term premium can predict substantial amounts of stock return variation. This phenomenon occurs over business cycle and longer horizons. Daily, weekly, and monthly stock returns are still close to unpredictable, and technical systems for predicting such movements are still close to useless.

Bond returns are predictable. Though the expectations model works well in the long run, a steeply upward sloping yield curve means that expected returns on long-term bonds are higher than on short-term bonds for the next year. These predictions are not guarantees—there is still substantial risk—but the tendency is discernible.

Foreign exchange returns are predictable. If you put your money in a country whose interest rates are higher than usual relative to the U.S., you expect to earn more money even after converting back to dollars. Again, this prediction is not a guarantee—exchange rates do vary, and a lot, so the strategy is risky.

Volatility does change through time. Times of past volatility indicate future volatility. Volatility also is higher after large price drops. Bond market volatility is higher when interest rates are higher, and possibly when interest rate spreads are higher as well.

3. Some mutual funds seem to outperform simple indexes, even after controlling for risk through market betas. Fund returns are also slightly predictable: Past winning funds seem to do better than average in the future, and past losing funds seem to do worse than average in the future. For a while, this seemed to indicate that there is some persistent skill in active management. However, multifactor models explain most fund persistence: Funds earn persistent returns by following fairly mechanical *styles*, not by persistent skill at stock selection.

Again, these statements are not dogma, but a cautious summary of a large body of careful empirical work. The strength and usefulness of many results are hotly debated, as are the underlying reasons for many of these new facts. But the old world is gone.

BOX 1

The CAPM and multifactor models

The CAPM uses a *time-series* regression to measure beta, β, which quantifies an asset's or portfolio's tendency to move with the market as a whole,

$$R_t^i - R_t^f = a_i + \beta_{im}(R_t^m - R_t^f) + \varepsilon_t^i; \quad t = 1, 2 \ldots T \text{ for each asset } i.$$

Then, the CAPM predicts that the expected excess return should be proportional to beta,

$$E(R_t^i - R_t^f) = \beta_{im}\lambda_m \text{ for each } i.$$

λ_m gives the "price of beta risk" or "market risk premium"—the amount by which expected returns must rise to compensate investors for higher beta. Since the model applies to the market return as well, we can measure λ_m via

$$\lambda_m = E(R_t^m - R_t^f).$$

Multifactor models extend this theory in a straightforward way. They use a time-series *multiple* regression to quantify an asset's tendency to move with multiple risk factors F^A, F^B, etc.

3) $$R_t^i - R_t^f = a_i + \beta_{im}(R_t^m - R_t^f) + \beta_{iA}F_t^A + \beta_{iB}F_t^B + \ldots + \varepsilon_t^i; \quad t = 1,2 \ldots T \text{ for each asset } i.$$

Then, the multifactor model predicts that the expected excess return is proportional to the betas

4) $$E(R_t^i - R_t^f) = \beta_{im}\lambda_m + \beta_{iA}\lambda_A + \beta_{iB}\lambda_B + \ldots \text{ for each } i.$$

The residual or unexplained average return in either case is called an alpha,

$$\alpha_i \equiv E(R_t^i - R_t^f) - (\beta_{im}\lambda_m + \beta_{iA}\lambda_A + \beta_{iB}\lambda_B + \ldots).$$

The CAPM and multifactor models

The CAPM

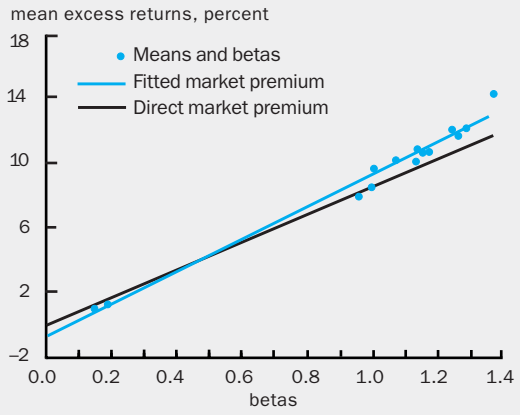
The CAPM proved stunningly successful in a quarter century of empirical work. Every strategy that seemed to give high average returns turned out to have a high beta, or a large tendency to move with the market. Strategies that one might have thought gave high average returns (such as holding very volatile stocks) turned out not to have high average returns when they did not have high betas.

Figure 1 presents a typical evaluation of the CAPM. I examine 10 portfolios of NYSE stocks sorted by size (total market capitalization), along with a portfolio of corporate bonds and long-term government bonds. As the vertical axis shows, there is a sizable spread in average returns between large stocks (lower average return) and small stocks (higher average return) and a large spread between stocks and bonds. The figure plots these average returns against market betas. You can see how the CAPM prediction fits: Portfolios with higher average returns have higher betas.

In fact, figure 1 captures one of the first significant *failures* of the CAPM. The smallest firms (the far right portfolio) seem to earn an average return a few percent too high given their betas. This is the celebrated "small-firm effect," (Banz, 1981) and this deviation is statistically significant. Would that all failed economic theories worked so well! However, the plot shows that this effect is within the range that statisticians can argue about. Estimating the slope of the line by fitting a cross-sectional regression (average return against beta), shown in the colored line, rather than forcing the line to go through the market and Treasury bill return, shown in the black line, halves

FIGURE 1

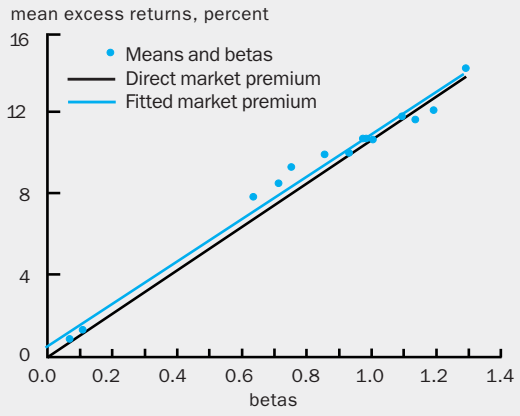
CAPM—Mean excess returns vs. beta, version 1



Notes: Average returns versus betas on the NYSE value-weighted portfolio for ten size-sorted stock portfolios, government bonds, and corporate bonds. Sample period 1947–96. The black line draws the CAPM prediction by fitting the market proxy and Treasury bill rates exactly (a time-series test) and the colored line draws the CAPM prediction by fitting an OLS cross-sectional regression to the displayed data points (a second-pass or cross-sectional test). The small-firm portfolios are at the top right. Moving down and to the left, one sees increasingly large-firm portfolios and the market index. The points far down and to the left are the government bond and Treasury bill returns.

FIGURE 2

CAPM—Mean excess returns vs. beta, version 2



Notes: CAPM using the equally weighted NYSE as the "market portfolio." Otherwise, this figure is identical to figure 1.

the small-firm effect. Figure 2 uses the equally weighted portfolio as market proxy, and this change in specification eliminates the small-firm effect, making the line of average returns versus betas if anything too shallow rather than too steep.

Why we expect multiple factors

In retrospect, it is surprising that the CAPM worked so well for so long. The assumptions on which it is built are very stylized and simplified. Asset pricing theory recognized at least since Merton (1973, 1971) the theoretical possibility, indeed probability, that we should need *factors, state variables* or *sources of priced risk*, beyond movements in the market portfolio to explain why some average returns are higher than others. (See box 1 for details of the CAPM and multifactor models.)

Most importantly, *the average investor has a job*. The CAPM (together with the use of the NYSE portfolio as the market proxy) simplifies matters by assuming that the average investor only cares about the performance of his investment portfolio. While there are investors like that, for most of us eventual wealth comes both from investment and from earning a living. Importantly, events like recessions hurt the majority of investors. Those who don't actually lose jobs get lower salaries or bonuses. A very limited number of people actually do better in a recession.

With this fact in mind, compare two stocks. They both have the same sensitivity to market movements. However, one of them does well in recessions, while the other does poorly. Clearly, most investors prefer the stock that does well in recessions, since its performance will cushion the blows to their other income. If lots of people feel that way, they bid up the price of that stock, or, equivalently, they are willing to hold it at a lower average return. Conversely, the procyclical stock's price will fall or it must offer a higher average return in order to get investors to hold it.

In sum, we should expect that procyclical stocks that do well in booms and worse in recessions will have to offer higher average returns than countercyclical stocks that do well in recessions, even if the stocks have the same market beta. We expect that *another dimension of risk*—covariation with recessions—will matter in determining average returns.[1]

What kinds of additional factors should we look for? Generally, asset pricing theory specifies that assets will have to pay high average returns if they do poorly in "bad times"—times in which investors would particularly like their investments not to perform badly and are willing to sacrifice some expected return in order to ensure that this is so. Consumption (or, more generally, marginal utility) should provide the purest measure of bad times. Investors consume less when their income prospects are low or if they think future returns will be bad. Low consumption thus *reveals* that this is indeed a time at which investors would especially like portfolios not to do badly, and would be willing to pay to ensure that wish. Alas, efforts to relate asset returns to consumption data are not (yet) a great success. Therefore, empirically useful asset pricing models examine more direct measures of good times or bad times. Broad categories of such indicators are

1. The market return. The CAPM is usually included and extended. People are unhappy if the market crashes.

2. Events, such as recessions, that drive investors' noninvestment sources of income.

3. Variables, such as the p/d ratio or slope of the yield curve, that forecast stock or bond returns (called "state variables for changing investment opportunity sets").

4. Returns on other well-diversified portfolios.

One formally justifies the first three factors by stating assumptions under which each variable is related to average consumption. For example, 1) if the market as a whole declines, consumers lose wealth and will cut back on consumption; 2) if a recession leads people to lose their jobs, then they will cut back on consumption; and, 3) if you are saving for retirement, then news that interest rates and average stock returns have declined is bad news, which will cause you to lower consumption. This last point establishes a connection between predictability of returns and the presence of additional risk factors for understanding

the cross-section of average returns. As pointed out by Merton (1971), one would give up some average return to have a portfolio that did well when there was bad news about future market returns.

The fourth kind of factor—additional portfolio returns—is most easily defended as a proxy for any of the other three. The fitted value of a regression of any pricing factor on the set of all asset returns is a portfolio that carries exactly the same pricing information as the original factor—a *factor-mimicking* portfolio.

It is vital that the extra risk factors affect the *average* investor. If an event makes investor A worse off and investor B better off, then investor A buys assets that do well when the event happens and investor B sells them. They transfer the risk of the event, but the price or expected return of the asset is unaffected. For a factor to affect prices or expected returns, it must affect the average investor, so investors collectively bid up or down the price and expected return of assets that covary with the event rather than just transferring the risk without affecting equilibrium prices.

Inspired by this broad direction, empirical researchers have found quite a number of specific factors that seem to explain the variation in average returns across assets. In general, empirical success varies inversely with theoretical purity.

Small and value/growth stocks

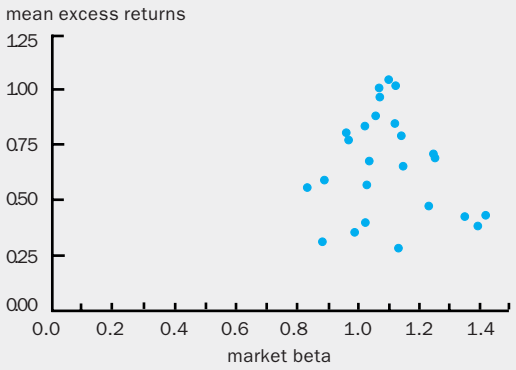
The size and book to market factors advocated by Fama and French (1996) are one of the most popular additional risk factors.

Small-cap stocks have small market values (price times shares outstanding). Value (or high book/market) stocks have market values that are small relative to the value of assets on the company's books. Both categories of stocks have quite high average returns. Large and growth stocks are the opposite of small and value and seem to have unusually low average returns. (See Fama and French, 1993, for a review.) The idea that low prices lead to high average returns is natural.

High average returns are consistent with the CAPM, if these categories of stocks have high sensitivity to the market, high betas. However, small and especially value stocks seem to have abnormally high returns even after accounting for market beta. Conversely, growth stocks seem to do systematically worse than their CAPM betas suggest. Figure 3 shows this value–size puzzle. It is just like figure 1, except that the stocks are sorted into portfolios based on size and book/market ratio[2] rather than size alone. The highest portfolios have *three* times the average excess return of the lowest portfolios, and this variation has nothing at all to do with market betas.

FIGURE 3

Mean excess returns vs. market beta, Fama–French portfolios



Notes: Average monthly returns versus market beta for 25 stock portfolios sorted on the basis of size and book/market ratio.

In figure 4, I connect portfolios of different sizes within the same book/market category (panel A). Variation in *size* produces a variation in average returns that is positively related to variation in market betas, as shown in figure 1. In panel B, I connect portfolios that have different book/market ratios within size categories. Variation in book/market ratio produces a variation in average return that is *negatively* related to market beta. Because of this value effect, the CAPM is a disaster when confronted with these portfolios.

To explain these facts, Fama and French (1993, 1996) advocate a multifactor model with the market return, the return of small less big stocks (SMB), and the return of high book/market less low book/market stocks (HML) as three factors. They show that variation in average returns of the 25 size and book/market portfolios can be explained by varying loadings (betas) on the latter two factors.

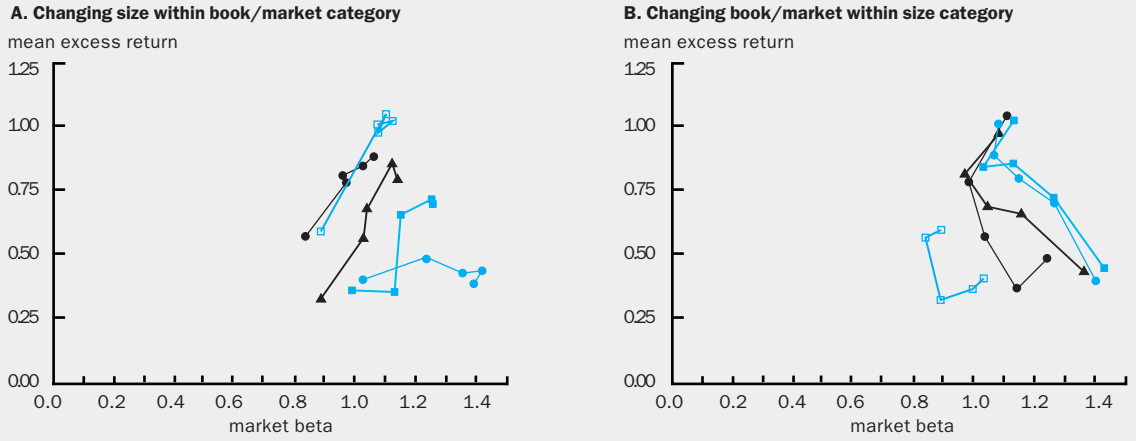
Figure 5 illustrates Fama and French's results. As in figure 4, the vertical axis is the average returns of the 25 size and book/market portfolios. Now, the horizontal axis is the predicted values from the Fama–French three-factor model. The points should all lie on a 45 degree line if the model is correct. The points lie much closer to this prediction in figure 5 than in figures 3 and 4. The worst fit is for the growth stocks (lowest line, panel A), for which there is little variation in average return despite large variation in size beta as one moves from small to large firms.

What are the size and value factors?

One would like to understand the real, macroeconomic, aggregate, nondiversifiable risk that is proxied by the returns of the HML and SMB portfolios. Why

FIGURE 4

Mean excess returns vs. market beta, varying size and book/market ratio



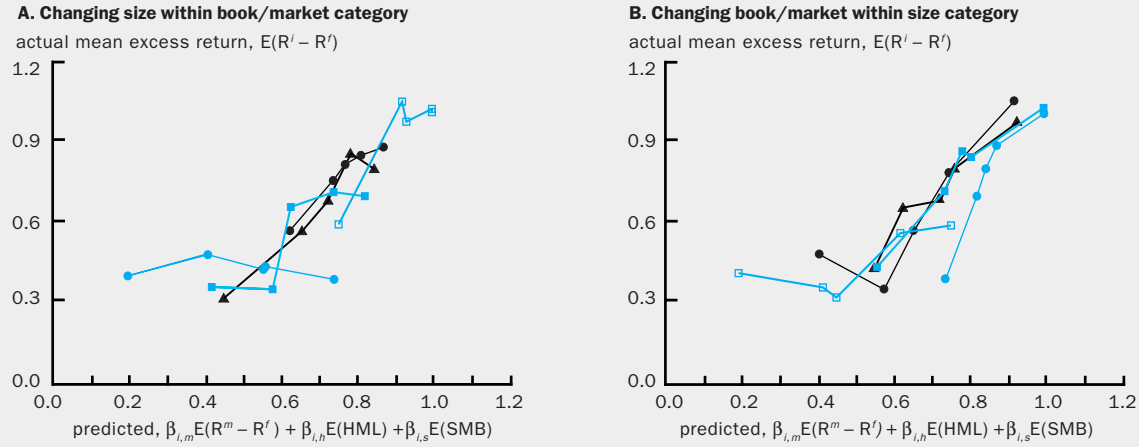
Notes: Average returns versus market beta for 25 stock portfolios sorted on the basis of size and book/market ratio. The points are the same as figure 3. In panel A, lines connect portfolios as size varies within book/market categories; in panel B, lines connect portfolios as book/market ratio varies within size categories.

are investors so concerned about holding stocks that do badly at the times that the HML (value less growth) and SMB (small-cap less large-cap) portfolios do badly, even though the market does not fall? The answer to this question is not yet totally clear.

Fama and French (1995) note that the typical value stock has a price that has been driven down due to financial distress. The stocks of firms on the verge of bankruptcy have recovered more often than not, which generates the high average returns of this strategy.[3] This observation suggests a natural interpretation of the value premium: In the event of a credit crunch, liquidity crunch, or flight to quality, stocks in financial distress will do very badly, and this is precisely when investors least want to hear that their portfolio is losing money. (One cannot count the "distress" of the individual firm as a risk factor. Such distress is idiosyncratic and can be diversified away. Only aggregate events that average investors care about can result in a risk premium.)

FIGURE 5

Mean excess return vs. three-factor model predictions



Notes: Average returns versus market beta for 25 stock portfolios sorted on the basis of size and book/market ratio versus predictions of Fama–French three-factor model. The predictions are derived by regressing each of the 25 portfolio returns, R^i_t, on the market portfolio, R^m_t, and the two Fama–French factor portfolios, SMB_t(small minus big) and HML_t(high minus low book/market). (See equation 4 in box 1.)

Heaton and Lucas's (1997) results add to this story for the value effect. They note that the typical stockholder is the proprietor of a small, privately held business. Such an investor's income is, of course, particularly sensitive to the kinds of financial events that cause distress among small firms and distressed value firms. Therefore, this investor would demand a substantial premium to hold value stocks and would hold growth stocks despite a low premium.

Liew and Vassalou (1999), among others, link value and small-firm returns to macroeconomic events. They find that in many countries, counterparts to HML and SMB contain supplementary information to that contained in the market return for forecasting gross domestic product (GDP) growth. For example, they report a regression

$$GDP_{t\to t+4} = a + 0.065\, MKT_{t-4\to t} + 0.058\, HML_{t-4\to t} + \varepsilon_{t+4},$$

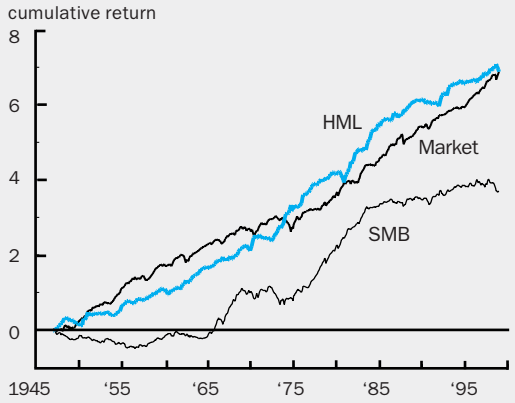
where $GDP_{t\to t+4}$ denotes the following year's GDP growth and $MKT_{t-4\to t}$ and $HML_{t-4\to t}$ denote the previous year's return on the market index and HML portfolio. Thus, a 10 percent HML return raises the GDP forecast by 0.5 percentage points. (Both coefficients are significant with t-statistics of 3.09 and 2.83, respectively.)

The effects are still under investigation. Figure 6 plots the cumulative return on the HML and SMB portfolios; a link between these returns and obvious macroeconomic events does not jump out. Both portfolios have essentially no correlation with the market return, though HML does seem to move inversely with large market declines. HML goes down more than the market in some business cycles, but less in others.

On the other hand, one can ignore Fama and French's motivation and regard the model as an *arbitrage pricing* theory (APT) following Ross (1976). If the returns of the 25 size and book/market portfolios could be *perfectly* replicated by the returns of the three-factor portfolios—if the R^2 values in the time-series regressions of the 25 portfolio on the three factors were 100 percent—then the multifactor model would have to hold exactly, in order to preclude arbitrage opportunities. To see this, suppose that one of the 25 portfolios—call it portfolio A—gives an average return 5 percent above the average return predicted by the Fama–French model, and its R^2 is 100 percent. Then, one could short a combination of the three-factor portfolios, buy portfolio A, and earn a completely riskless profit. This logic is often used to argue that a *high* R^2 should imply an *approximate* multifactor model. If the R^2 were only 95 percent, then an average return 5 percent above the factor model prediction

FIGURE 6

Cumulative returns on market portfolios



Notes: Cumulative returns on the market RMRF, SMB, and HML portfolios. The SMB return is formed by $R_t^{rs} + aSMB_t$; $a = \sigma(RMRF)/\sigma(SMB)$. In this way it is a return that can be cumulated rather than a zero-cost portfolio, and its standard deviation is equal to that of the market return. HML is adjusted similarly. The vertical axis is the log base 2 of the cumulative return or value of $1 invested at the beginning of the sample period. Thus, each time a line increases by 1 unit, the value doubles.

would imply that the strategy long portfolio A and short a combination of the three-factor portfolios would earn a very high average return with very little, though not zero, risk—a very high Sharpe ratio.

In fact, the R^2 values of Fama and French's (1993) time-series regressions are all in the 90 percent to 95 percent range, so extremely high risk prices for the residuals would have to be invoked for the model *not* to fit well. Conversely, given the average returns from HML and SMB and the failure of the CAPM to explain those returns, there would be near-arbitrage opportunities if value and small stocks did not move together in the way described by the Fama–French model.

One way to assess whether the three factors proxy for real macroeconomic risks is by checking whether the multifactor model prices additional portfolios, especially portfolios whose ex-post returns are not well explained by the factors (portfolios that do not have high R^2 values in time-series regressions). Fama and French (1996) find that the SMB and HML portfolios comfortably explain strategies based on alternative price multiples (price/earnings, book/market), five-year sales growth (this is the only strategy that does not form portfolios based on price variables), and the tendency of five-year returns to reverse. All of these strategies are not explained by CAPM betas. However, they all also produce portfolios with high R^2 values in a time-series regression on the HML and SMB portfolios. This is good and bad news. It might

mean that the model is a good APT, and that the size and book/market characteristics describe the major sources of priced variation in all stocks. On the other hand, it might mean that these extra ways of constructing portfolios just haven't identified other sources of priced variation in stock returns. (Fama and French, 1996, also find that HML and SMB do not explain *momentum*, despite high R^2 values. I discuss this anomaly below.) The portfolios of stocks sorted by industry in Fama and French (1997) have lower R^2 values, and the model works less well.

A final concern is that the size and book/market premiums seem to have diminished substantially in recent years. The sharp decline in the SMB portfolio return around 1980 when the small-firm effect was first popularized is obvious in figure 6. In Fama and French's (1993) initial samples, 1960–90, the HML cumulative return starts about one-half (0.62) below the market and ends up about one-half (0.77) above the market. On the log scale of the figure, this corresponds to Fama and French's report that the HML average return is about double (precisely, $2^{0.62+0.77} = 2.6$ times) that of the market. However, over the entire sample of the plot, the HML portfolio starts and ends at the same place and so earns almost exactly the same as the market. From 1990 to now, the HML portfolio loses about one-half relative to the market, meaning an investor in the market has increased his money one and a half times as much as an HML investor. (The actual number is 0.77 so the market return is $2^{0.77} = 1.71$ times better than the HML return.)

Among other worries, if the average returns decline right after publication it suggests that the anomalies may simply have been overlooked by a large fraction of investors. As they move in, prices go up further, helping the apparent anomaly for a while. But once a large number of investors have moved in to include small and value stocks in their portfolios, the anomalous high average returns disappear.

However, average returns are hard to measure. There have been previous ten- to 20-year periods in which small stocks did very badly, for example the 1950s, and similar decade-long variations in the HML premium. Also, since SMB and HML have a beta of essentially zero on the market, *any* upward trend is a violation of the CAPM and says that investors can improve their overall mean–variance tradeoff by taking on some of the HML or SMB portfolio.

Macroeconomic factors

I focus on the size and value factors because they provide the most empirically successful multifactor model and have attracted much industry as well as academic attention. Several authors have used macroeconomic variables as factors. This procedure examines directly whether stock performance during bad macroeconomic times determines average returns. Jagannathan and Wang (1996) and Reyfman (1997) use labor income; Chen, Roll, and Ross (1986) look at industrial production and inflation among other variables; and Cochrane (1996) looks at investment growth. All these authors find that average returns line up with betas calculated using the macroeconomic indicators. The factors are theoretically easier to motivate, but none explains the value and size portfolios as well as the (theoretically less solid, so far) size and value factors.

Merton's (1973, 1971) theory says that variables which predict market returns should show up as factors that explain cross-sectional variation in average returns. Campbell (1996) is the lone test I know of to directly address this question. Cochrane (1996) and Jagannathan and Wang (1996) perform related tests in that they include "scaled return" factors, for example, market return at t multiplied by d/p ratio at $t - 1$; they find that these factors are also important in understanding cross-sectional variation in average returns.

The next step is to link these more fundamentally determined factors with the empirically more successful value and small-firm factor portfolios. Because of measurement difficulties and selection biases, fundamentally determined macroeconomic factors will never approach the empirical performance of portfolio-based factors. However, they may help to explain which portfolio-based factors really work and why.

Predictable returns

The view that risky asset returns are largely unpredictable, or that prices follow "random walks," remains immensely successful (Malkiel, 1990, is a classic and readable introduction). It is also widely ignored.

Unpredictable returns mean that if stocks went up yesterday, there is no exploitable tendency for them to decline today because of "profit taking" or to continue to rise today because of "momentum." "Technical" signals, including analysis of past price movements trading volume, open interest, and so on are close to useless for forecasting short-term gains and losses. As I write, value funds are reportedly suffering large outflows because their stocks have done poorly in the last few months, leading fund investors to move money into blue-chip funds that have performed better (New York Times Company, 1999). Unpredictable returns mean that this strategy will not do anything for investors' portfolios over the long run except rack up trading costs. If funds are selling stocks, then

contrarian investors must be buying them, but unpredictable returns mean that this strategy can not improve performance either. If one can not systematically make money, one can not systematically lose money either.

As discussed in the introduction, researchers once believed that stock returns (more precisely, the excess returns on stocks over short-term interest rates) were completely unpredictable. It now turns out that average returns on the market and individual securities *do* vary over time and that stock returns *are* predictable. Alas for would-be technical traders, much of that predictability comes at long horizons and seems to be associated with business cycles and financial distress.

Market returns

Table 1 presents a regression that forecasts returns. Low prices—relative to dividends, book value, earnings, sales, or other divisors—predict higher subsequent returns. As the R^2 values in table 1 show, these are long-horizon effects: Annual returns are only slightly predictable and month-to-month returns are still strikingly unpredictable, but returns at five-year horizons seem very predictable. (Fama and French, 1989, is an excellent reference for this kind of regression).

The results at different horizons are reflections of a single underlying phenomenon. If daily returns are very slightly predictable by a slow-moving variable, that predictability adds up over long horizons. For example, you can predict that the temperature in Chicago will rise about one-third of a degree per day in spring. This forecast explains very little of the day to day variation in temperature, but tracks almost all of the rise in temperature from January to July. Thus, the R^2 rises with horizon.

Precisely, suppose that we forecast returns with a forecasting variable x, according to

$$1) \quad R_{t+1} - R^{TB}_{t+1} = a + bx_t + \varepsilon_{t+1}$$

$$2) \quad x_{t+1} = c + \rho x_t + \delta_{t+1}.$$

Small values of b and R^2 in equation 1 and a large coefficient ρ in equation 2 imply mathematically that the long-horizon regression as in table 1 has a large regression coefficient b and large R^2.

This regression has a powerful implication: Stocks are in many ways like bonds. Any bond investor understands that a string of good past returns that pushes the price up is bad news for subsequent returns. Many stock investors see a string of good past returns and become elated that we seem to be in a "bull market," concluding future stock returns will be good as well. The regression reveals the opposite: A string of good past returns which drives up stock prices is bad news for subsequent stock returns, as it is for bonds.

Long-horizon return predictability was first documented in the volatility tests of Shiller (1981) and LeRoy and Porter (1981). They found that stock prices vary far too much to be accounted for by changing expectations of subsequent cash flows; thus changing discount rates or expected returns must account for variation in stock prices. These volatility tests turn out to be almost identical to regressions such as those in table 1 (Cochrane, 1991).

Momentum and reversal

Since a string of good returns gives a high price, it is not surprising that individual stocks that do well for a long time (and reach a high price) subsequently do poorly, and stocks that do poorly for a long time (and reach a low price, market value, or market to book ratio) subsequently do well. Table 2, taken from Fama and French (1996) confirms this hunch. (Also, see DeBont and Thaler, 1985, and Jegadeesh and Titman, 1993.)

The first row in table 2 tracks the average monthly return from the *reversal* strategy. Each month, allocate all stocks to ten portfolios based on performance from year –5 to year –1. Then, buy the best-performing portfolio and short the worst-performing portfolio. This strategy earns a hefty –0.74 percent monthly return.[4] Past long-term losers come back and past winners do badly. Fama and French (1996) verify that these portfolio returns are explained by their three-factor model. Past winners move with value stocks,

TABLE 1

OLS regression of excess returns on price/dividend ratio

Horizon *k*	*b*	Standard error	R^2
1 year	–1.04	0.33	0.17
2 years	–2.04	0.66	0.26
3 years	–2.84	0.88	0.38
5 years	–6.22	1.24	0.59

Notes: OLS regressions of excess returns (value-weighted NYSE–Treasury bill rate) on value-weighted price/dividend ratio.

$$R^{VW}_{t\to t+k} - R^{TB}_{t\to t+k} = a + b(P_t / D_t) + \varepsilon_{t+k}.$$

$R_{t\to t+k}$ indicates the k year return. Standard errors use GMM to correct for heteroskedasticity and serial correlation.

TABLE 2

Average monthly returns, reversal and momentum strategies

Strategy	Period	Portfolio formation (months)	Average return, 10–1 (monthly %)
Reversal	July 1963–Dec. 1993	60–13	–0.74
Momentum	July 1963–Dec. 1993	12–2	+1.31
Reversal	Jan. 1931–Feb. 1963	60–13	–1.61
Momentum	Jan. 1931–Feb. 1963	12–2	+0.38

Notes: Each month, allocate all NYSE firms to 10 portfolios based on their performance during the "portfolio formation months" interval. For example, 60–13 forms portfolios based on returns from 5 years ago to 1 year, 1 month ago. Then buy the best-performing decile portfolio and short the worst-performing decile portfolio.

Source: Fama and French (1996, table 6).

and so inherit the value stock premium. (To compare the strategies, the table always buys the winners and shorts the losers. In practice, of course, you buy the losers and short the winners to earn +0.71 percent monthly average return.)

The second row of table 2 tracks the average monthly return from a *momentum* strategy. Each month, allocate all stocks to ten portfolios based on performance in the last *year*. Now, the winners continue to win and the losers continue to lose, so that buying the winners and shorting the losers generates a positive 1.31 percent monthly return.

Momentum is not explained by the Fama–French (1996) three-factor model. The past losers have low prices and tend to move with value stocks. Hence, the model predicts that they should have high average returns, not low average returns.

Momentum stocks move together, as do value and small stocks, so a "momentum factor" works to "explain" momentum portfolio returns (Carhart, 1997). This step is so obviously ad hoc (that is, an APT factor that will only explain returns of portfolios organized on the same characteristic as the factor rather than a proxy for macroeconomic risk) that most people are uncomfortable adding it. We obviously do not want to add a new factor for every anomaly.

Is momentum really there, and if so, is it exploitable after transaction costs? One warning is that it does not seem stable over subsamples. The third and fourth lines in table 2 show that the momentum effect essentially disappears in the earlier data sample, while reversal is even stronger in that sample.

Momentum is really just a new way of looking at an old phenomenon, the small apparent predictability of monthly individual stock returns. A tiny regression R^2 for forecasting monthly returns of 0.0025 (0.25 percent) is more than adequate to generate the momentum results of table 2. The key is the large standard deviation of individual stock returns, typically 40 percent or more on an annual basis. The average return of the best performing decile of a normal distribution is 1.76 standard deviations above the mean,[5] so the winning momentum portfolio went up about 80 percent in the previous year and the typical losing portfolio went down about 60 percent. Only a small amount of continuation will give a 1 percent monthly return when multiplied by such large past returns. To be precise, the monthly individual stock standard deviation is about $40\% / \sqrt{12} \approx 12\%$. If the R^2 is 0.0025, the standard deviation of the predictable part of returns is $\sqrt{.0025} \times 12\% \approx 0.6\%$. Hence, the decile predicted to perform best will earn $1.76 \times 0.6\% \approx 1\%$ above the mean. Since the strategy buys the winners and shorts the losers, an R^2 of 0.0025 implies that one should earn a 2 percent monthly return by the momentum strategy.

We have known at least since Fama (1965) that monthly and higher frequency stock returns have slight, statistically significant predictability with R^2 about 0.01. Campbell, Lo, and MacKinlay (1997, table 2.4) provide an updated summary of index autocorrelations (the R^2 is the squared autocorrelation), part of which I show in table 3. Note the correlation of the equally weighted portfolio, which emphasizes small stocks.[6]

However, such small, though statistically significant, high-frequency predictability has thus far failed to yield exploitable profits after one takes into account transaction costs, thin trading of small stocks, and high short-sale costs. The momentum strategy for exploiting this correlation may not work in practice for the same reasons. Momentum does require frequent trading. The portfolios in table 2 are re-formed every

TABLE 3

First-order autocorrelation, CRSP value- and equally weighted index returns

Frequency	Portfolio	Correlation ρ_1
Daily	Value-weighted	0.18
	Equally weighted	0.35
Monthly	Value-weighted	0.043
	Equally weighted	0.17

Note: Sample 1962–94.
Source: Campbell, Lo, and MacKinlay (1997).

month. Annual winners and losers will not change that often, but the winning and losing portfolio must be turned over at least once per year. In a quantitative examination of this effect, Carhart (1997) concludes that momentum is not exploitable after transaction costs are taken into account. Moskowitz and Grinblatt (1999) note that most of the apparent gains from the momentum strategy come from short positions in small illiquid stocks. They also find that a large part of momentum profits come from short positions taken in November. Many investors sell losing stocks toward the end of December to establish tax losses. By shorting illiquid losing stocks in November, an investor can profit from the selling pressure in December. This is also an anomaly, but it seems like a glitch rather than a central principle of risk and return in asset markets.

Even if momentum and reversal are real and as strong as indicated by table 2, they do not justify much of the trading based on past results that many investors seem to do. To get the 1 percent per month momentum return, one buys a portfolio that has typically gone up 80 percent in the last year, and shorts a portfolio that has typically gone down 60 percent. Trading between stocks and fund categories such as value and blue-chip with smaller past returns yields at best proportionally smaller results. Since much of the momentum return seems to come from shorting small illiquid stocks, mild momentum strategies may yield even less. And we have not quantified the substantial risk of momentum strategies.

Bonds

The venerable expectations model of the term structure specifies that long-term bond yields are equal to the average of expected future short-term bond yields (see box 2). For example, if long-term bond yields are higher than short-term bond yields—if the yield curve is upward sloping—this means that short-term rates are expected to rise in the future. The rise in future short-term rates means that investors can expect

BOX 2

Bond definitions and expectations hypothesis

Let $p_t^{(N)}$ denote the log of the N year discount bond price at time t. The N period continuously compounded yield is defined by $y_t^{(N)} = -\frac{1}{N} p_t^{(N)}$. The continuously compounded holding period return is the selling price less the buying price, $hpr_{t+1}^{(N)} = p_{t+1}^{(N-1)} - p_t^{(N)}$. The forward rate is the rate at which an investor can contract today to borrow money $N-1$ years from now, and repay that money N years from now. Since an investor can synthesize a forward contract from discount bonds, the forward rate is determined from discount bond prices by

$$f_t^{(N)} = p_t^{(N-1)} - p_t^{(N)}.$$

The "spot rate" refers, by contrast with a forward rate, to the yield on any bond for which the investor take immediate delivery. Forward rates are typically higher than spot rates when the yield curve rises, since the yield is the average of intervening forward rates,

$$y_t^{(N)} = \frac{1}{N}(f_t^{(1)} + f_t^{(2)} + f_t^{(3)} + \ldots + f_t^{(N)}).$$

The expectations hypothesis states that the expected log or continuously compounded return should be the same for any bond strategy. This statement has three mathematically equivalent expressions:

1. The forward rate should equal the expected value of the future spot rate,

$$f_t^{(N)} = E_t(y_{t+N-1}^{(1)}).$$

2. The expected holding period return should be the same on bonds of any maturity

$$E_t(hpr_{t+1}^{(N)}) = E_t(hpr_{t+1}^{(M)}) = y_t^{(1)}.$$

3. The long-term bond yield should equal the average of the expected future short rates,

$$y_t^{(N)} = \frac{1}{N} E_t(y_t^{(1)} + y_{t+1}^{(1)} + \ldots + y_{t+N-1}^{(1)}).$$

The expectations hypothesis is often amended to allow a constant risk premium of undetermined sign in these equations. Its violation is then often described as evidence for a "time-varying risk premium."

The expectations hypothesis is not quite the same thing as risk-neutrality, because the expected log return is not equal to the log expected return. However, the issues here are larger than the difference between the expectations hypothesis and strict risk-neutrality.

TABLE 4

Zero-coupon bond returns

Maturity N	Average holding period return	Standard error	Standard deviation
1	5.83	0.42	2.83
2	6.15	0.54	3.65
3	6.40	0.69	4.66
4	6.40	0.85	5.71
5	6.36	0.98	6.58

Note: Continuously compounded one-year holding period returns on zero-coupon bonds of varying maturity. Annual data from CRSP 1953–97.

the same rate of return whether they hold a long-term bond to maturity or roll over short-term bonds with initially low returns and subsequent higher returns.

As with the CAPM and the view that stock returns are independent over time, a new round of research has significantly modified this traditional view of bond markets.

Table 4 calculates the average return on bonds of different maturities. The expectations hypothesis seems to do pretty well. Average holding period returns do not seem very different across bond maturities, despite the increasing standard deviation of longer-maturity bond returns. The small increase in average returns for long-term bonds, equivalent to a slight average upward slope in the yield curve, is usually excused as a "liquidity premium." Table 4 is just the tip of an iceberg of successes for the expectations model. Especially in times of significant inflation and exchange rate instability, the expectations hypothesis has done a very good first-order job of explaining the term structure of interest rates.

However, if there are times when long-term bonds are expected to do better and other times when short-term bonds are expected to do better, the unconditional averages in table 4 could still show no pattern. Similarly, one might want to check whether a forward rate that is *unusually high* forecasts an unusual *increase* in spot rates.

Table 5 updates Fama and Bliss's (1987) classic regression tests of this idea. Panel A presents a regression of the change in yields on the forward-spot spread. (The forward-spot spread measures the slope of the yield curve.) The expectations hypothesis predicts a slope coefficient of 1.0, since the forward rate should equal the expected future spot rate. If, for example, forward rates are lower than expected future spot rates, traders can lock in a borrowing position with a forward contract and then lend at the higher spot rate when the time comes.

Instead, at a one-year horizon we find slope coefficients near zero and a negative adjusted R^2. Forward rates one year out seem to have no predictive power whatsoever for changes in the spot rate one year from now. On the other hand, by four years out, we see slope coefficients within one standard error of 1.0. Thus, the expectations hypothesis seems to do poorly at short (one-year) horizons, but much better at longer horizons.

If the expectations hypothesis does not work at one-year horizons, then there is money to be made—one must be able to foresee years in which short-term bonds will return more than long-term bonds and vice versa, at least to some extent. To confirm this implication, panel B of table 5 runs regressions of the one-year excess return on long-term bonds on the forward-spot spread. Here, the expectations hypothesis predicts a coefficient of zero: No signal (including the

TABLE 5

Forecasts based on forward-spot spread

	A. Change in yields					B. Holding period returns				
N	Intercept	Standard error, intercept	Slope	Standard error, slope	Adjusted R^2	Intercept	Standard error, intercept	Slope	Standard error, slope	Adjusted R^2
1	0.10	0.3	–0.10	0.36	–0.020	–0.1	0.3	1.10	0.36	0.16
2	–0.01	0.4	0.37	0.33	0.005	–0.5	0.5	1.46	0.44	0.19
3	–0.04	0.5	0.41	0.33	0.013	–0.4	0.8	1.30	0.54	0.10
4	–0.30	0.5	0.77	0.31	0.110	–0.5	1.0	1.31	0.63	0.07

Notes: OLS regressions, 1953–97 annual data. Panel A estimates the regression $y_{t+n}^{(1)} - y_t^{(1)} = a + b\,(f_t^{(N+1)} - y_t^{(1)}) + \varepsilon_{t+N}$ and panel B estimates the regression $hpr_{t+1}^{(N)} - y_t^{(1)} = a + b\,(f_t^{(N+1)} - y_t^{(1)}) + \varepsilon_{t+1}$, where $y_t^{(N)}$ denotes the N-year bond yield at date t; $f_t^{(N)}$ denotes the N-period ahead forward rate; and $hpr_{t+1}^{(N)}$ denotes the one-year holding period return at date $t+1$ on an N-year bond. Yields and returns in annual percentages.

forward-spot spread) should be able to tell you that this is a particularly good time for long bonds versus short bonds, as the random walk view of stock prices says that no signal should be able to tell you that this is a particularly good or bad day for stocks versus bonds. However, the coefficients in panel B are all about 1.0. A high forward rate does not indicate that interest rates will be higher one year from now; it seems to indicate that investors will earn that much more by holding long-term bonds.[7]

Of course, there is risk. The R^2 values are all 0.1–0.2, about the same values as the R^2 from the d/p regression at a one-year horizon, so this strategy will often go wrong. Still, 0.1–0.2 is not zero, so the strategy does pay off more often than not, in violation of the expectations hypothesis. Furthermore, the forward-spot spread is a slow-moving variable, typically reversing sign once per business cycle. Thus, the R^2 builds with horizon as with the d/p regression, peaking in the 30 percent range (Fama and French, 1989).

Foreign exchange

Suppose interest rates are higher in Germany than in the U.S. Does this mean that one can earn more money by investing in German bonds? There are several reasons that the answer might be no. First, of course, is default risk. Governments have defaulted on bonds in the past and may do so again. Second, and more important, is the risk of devaluation. If German interest rates are 10 percent and U.S. interest rates are 5 percent, but the euro falls 5 percent relative to the dollar during the year, you make no more money holding the German bonds despite their attractive interest rate. Since lots of investors are making this calculation, it is natural to conclude that an interest rate differential across countries on bonds of similar credit risk should reveal an expectation of currency devaluation. The logic is exactly the same as that of the expectations hypothesis in the term structure. Initially attractive yield or interest rate differentials should be met by an offsetting event so that you make no more money on average in one maturity or currency versus another.[8]

As with the expectations hypothesis in the term structure, the expected depreciation view still constitutes an important first-order understanding of interest rate differentials and exchange rates. For example, interest rates in east Asian currencies were very high on the eve of the recent currency tumbles, and many banks were making tidy sums borrowing at 5 percent in dollars to lend at 20 percent in local currencies. This suggests that traders were anticipating a 15 percent devaluation, or a smaller chance of a larger devaluation, which is exactly what happened. Many observers attribute high nominal interest rates in troubled economies to "tight monetary policy" aimed at defending the currency. In reality, high nominal rates reflect a large probability of inflation and devaluation—loose monetary policy—and correspond to much lower real rates.

Still, does a 5 percent interest rate differential correspond to a 5 percent expected depreciation, or does some of it represent a high expected return from holding debt in that country's currency? Furthermore, while expected depreciation is clearly a large part of the interest rate story in high-inflation economies, how does the story play out in economies like the U.S. and Germany, where inflation rates diverge little but exchange rates still fluctuate a large amount?

The first row of table 6 (from Hodrick, 2000, and Engel, 1996) shows the average appreciation of the dollar against the indicated currency over the sample period. The dollar fell against the deutschemark, yen, and Swiss franc, but appreciated against the pound sterling. The second row gives the average interest rate differential—the amount by which the foreign interest rate exceeds the U.S. interest rate.[9] According to the expectations hypothesis, these two numbers should be equal—interest rates should be higher in countries whose currencies depreciate against the dollar.

TABLE 6

Forward discount puzzle

	Deutsche-mark	Pound sterling	Yen	Swiss franc
Mean appreciation	–1.8	3.6	–5.0	–3.0
Mean interest differential	–3.9	2.1	–3.7	–5.9
b, 1975–89	–3.1	–2.0	–2.1	–2.6
R^2	.026	.033	.034	.033
b, 1976–96	–0.7	–1.8	–2.4	–1.3
b, 10-year horizon	0.8	0.6	0.5	–

Notes: The first row gives the average appreciation of the dollar against the indicated currency, in percent per year. The second row gives the average interest differential—foreign interest rate less domestic interest rate, measured as the forward premium—the 30-day forward rate less the spot exchange rate. The third through sixth rows give the coefficients and R^2 in a regression of exchange rate changes on the interest differential,

$$s_{t+1} - s_t = a + b\,(r_t^f - r_t^d) + \varepsilon_{t+1},$$

where s = log spot exchange rate, r^f = foreign interest rate, and r^d = domestic interest rate.

Source: Hodrick (2000), Engel (1996), and Meredith and Chinn (1998).

The second row shows roughly the expected pattern. Countries with steady long-term inflation have steadily higher interest rates and steady depreciation. The numbers in the first and second rows are not exactly the same, but exchange rates are notoriously volatile so these averages are not well measured. Hodrick (2000) shows that the difference between the first and second rows is not statistically different from zero. This fact is analogous to the evidence in table 4 that the expectations hypothesis works well *on average* for U.S. bonds.

As in the case of bonds, however, we can ask whether times of *temporarily* higher or lower interest rate differentials correspond to times of above- and below-average depreciation as they should. The third and fifth rows of table 6 update Fama's (1984) regression tests. The number here should be +1.0 in each case—1 percentage point extra interest differential should correspond to 1 percentage point extra expected depreciation. On the contrary, as table 6 shows, a higher than usual interest rate abroad seems to lead to further *appreciation.* This is the *forward discount puzzle*. See Engel (1996) and Lewis (1995) for recent surveys of the avalanche of academic work investigating whether this puzzle is really there and why.

The R^2 values shown in table 6 are quite low. However, like d/p and the term spread, the interest differential is a slow-moving forecasting variable, so the return forecast R^2 builds with horizon. Bekaert and Hodrick (1992) report that the R^2 rises to the 30 percent to 40 percent range at six-month horizons and then declines. That's high, but not 100 percent; taking advantage of any predictability strategy is quite risky.

The puzzle does *not* say that one earns more by holding bonds from countries with higher interest rates than others. Average inflation, depreciation, and interest rate differentials line up as they should. The puzzle *does* say that one earns more by holding bonds from countries whose interest rates are *higher than usual* relative to U.S. interest rates (and vice versa). The fact that the "usual" rate of depreciation and interest differential changes through time will, of course, diminish the out-of-sample performance of these trading rules.

One might expect that exchange rate depreciation works better for long-run exchange rates, as the expectations hypothesis works better for long-run interest rate changes. The last row of table 6, taken from Meredith and Chinn (1998) verifies that this is so. Ten-year exchange rate changes are correctly forecast by the interest differentials of ten-year bonds.

Mutual funds

Studying the returns of funds that follow a specific strategy gives us a way to assess whether that strategy works in practice, after transaction costs and other trading realities are taken into account. Studying the returns of actively managed funds tells us whether the time, talent, and effort put into picking securities pays off. Most of the literature on evaluating fund performance is devoted to the latter question.

A large body of empirical work, starting with Jensen (1969), finds that actively managed funds, on average, underperform the market index. I use data from Carhart (1997), whose measures of fund performance account for *survivor bias*. Survivor bias arises because funds that do badly go out of business. Therefore, the average fund that is alive at any point in time has an artificially good track record.

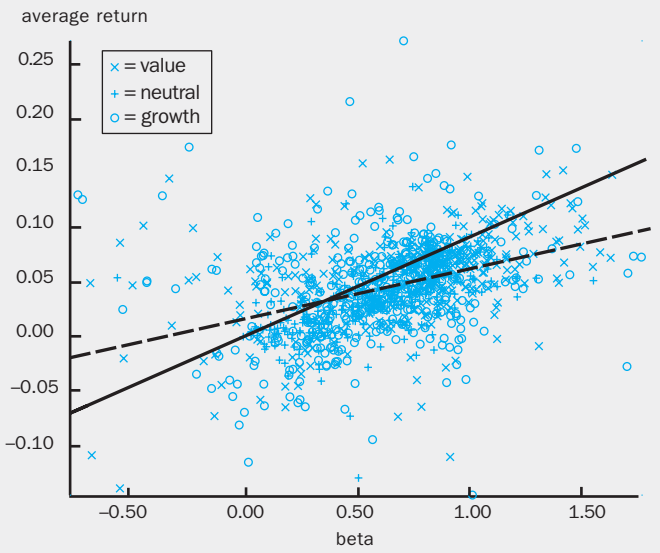
As with the stock portfolios in figure 1, the fund data in figure 7 show a definite correlation between beta and average return: Funds that did well took on more market risks. A cross-sectional regression line is a bit flatter than the line drawn through the Treasury bill and market return, but this is a typical result of measurement error in the betas. (The data are annual, and many funds are only around for a few years, contributing to beta measurement error.) The average fund underperforms the line connecting Treasury bills and the market index by 1.23 percent per year (that is, the average alpha is –1.23 percent).

The wide dispersion in fund average returns in figure 7 is a bit surprising. Average returns vary across funds almost as much as they do across individual stocks. This fact implies that the majority of funds are *not* holding well-diversified portfolios that would reduce return variation, but rather are loading up on specific bets.

Initially, the fact that the *average* fund underperforms the market seems beside the point. Perhaps the average fund is bad, but we want to know whether the good funds are any good. The trouble is, we must somehow distinguish skill from luck. The only way to separate skill from luck is to group funds based on some ex-ante observable characteristic, and then examine the average performance of the group. Of course, skillful funds should have done better, on average, in the past, and should continue to do better in the future. Thus, if there is skill in stock picking, we should see some persistence in fund performance. However, a generation of empirical work found no persistence at all. Funds that did well in the past were no more likely to do well in the future.

FIGURE 7

Average returns of mutual funds vs. market betas



Notes: Average returns of mutual funds over the Treasury bill rate versus their market betas. Sample consists of all funds with average total net assets greater than $25 million and more than 25 percent of their assets in stocks, in the Carhart (1996) database. Data sample 1962–96. The average excess return is computed as $E(R^i - R^f) = a_i + \beta_i \times 9\%$. a_i and β_i are computed from a time-series regression of fund annual excess returns on market annual excess returns over the life of the fund. The o, +, and x labels in the figure sort funds into thirds based on their regression coefficient h on the Fama–French value (HML) portfolio. The breakpoints are $h = -0.084$, 0.34. The dashed line gives the fit of a cross-sectional ordinary least squares regression of a_i on β_i; The solid line connects the Treasury bill (β and excess return = 0) and the market return ($\beta = 1$, excess return = 9%).

Since the average fund underperforms the market, and fund returns are not predictable, we conclude that active management does not generate superior performance, especially after transaction costs and fees. This fact is surprising. Professionals in almost any field do better than amateurs. One would expect that a trained experienced professional who spends all day reading about markets and stocks should be able to outperform simple indexing strategies. Even if entry into the industry is so easy that the *average* fund does not outperform simple indexes one would expect a few stars to outperform year after year, as good teams win championship after championship. Alas, the contrary fact is the result of practically every investigation, and even the anomalous results document very small effects.

Funds and value

Given the value, small-firm, and predictability effects, the idea that funds cluster around the market line is quite surprising. All of these new facts imply inescapably that there are simple, mechanical strategies that can give a risk/reward ratio greater than that of buying and holding the market index. Fama and French (1993) report that the HML portfolio alone gives nearly double the market Sharpe ratio—the same average return at half the standard deviation. Why don't funds cluster around a risk/reward line significantly above the market's?

Of course, we should not expect *all* funds to cluster around a higher risk/reward tradeoff. The average investor holds the market, and if funds are large enough, so must the average fund. Index funds, of course, will perform like the index. Still, the typical actively managed fund advertises high mean and, perhaps, low variance. No fund advertises cutting average returns in half to spare investors exposure to nonmarket sources of risk. Such funds, apparently aimed at mean–variance investors, should cluster around the highest risk/reward tradeoff available from mechanical strategies (and more, if active management does any good). Most troubling, funds who *say* they follow value strategies don't outperform the market either. For example, Lakonishok, Shleifer, and Vishny (1992, table 3) find that the average value fund underperforms the S&P500 by 1 percent just like all the others.

We can resolve this contradiction if we think that fund managers were simply unaware of the possibilities offered by our new facts, and so (despite the advertising) were not really following them. That seems to be the implication of figure 7, which sorts funds by their HML beta. One would expect the high-HML beta funds to outperform the market line. But the cutoff for the top one-third of funds is only a HML beta of 0.3, and even that may be high (many funds don't last long, so betas are poorly measured; the distribution of measured betas is wider than the actual distribution). Thus, the "value funds" were really not following the "value strategy" that earns the HML returns; if they were doing so they would have HML betas of 1.0. Similarly, Lakonishok, Shleifer, and Vishny's (1992) documentation of value funds' underperformance reveals that their market beta is close to 1.0. These results imply that value funds are not really following a value strategy, since their returns correlate with the market portfolio and not the value portfolio.

Interestingly, the number of value and small-cap funds (as revealed by their betas, not their marketing claims) is increasing quickly. Before 1990, 14 percent of funds had measured SMB betas greater than 1.0,

and 12 percent had HML betas greater than 1.0. In the full sample, both numbers have *doubled* to 22 percent and 23 percent. This trend suggests that funds will, in the future, be much less well described by the market index.

The view that funds were unaware of value strategies, and are now moving quickly to exploit them, can explain why most funds still earn near the market return, rather than the higher value return. However, this view contradicts the view that the value premium is an equilibrium risk premium, that is, that everyone knew about the value returns but chose not to invest all along because they feared the risks of value strategies. If it is not an equilibrium risk premium, it won't last long.

Persistence in fund returns

The fund counterpart to momentum in stock returns has been more extensively investigated than the value and size effects. Fund returns have also been found to be persistent. Since such persistence can be interpreted as evidence for persistent skill in picking stocks, it is not surprising that it has attracted a great deal of attention, starting with Hendricks, Patel, and Zeckhauser (1993).

Table 7, taken from Carhart (1997), shows that a portfolio of the best-performing one-thirtieth of funds last year outperforms a portfolio of the worst-performing one-thirtieth of funds by 1 percent per month (column 2). This is about the same size as the momentum effect in stocks, and similarly results from a small autocorrelation plus a large standard deviation in individual fund returns. This result verifies that mutual fund performance is persistent.

Perhaps the funds that did well took on more market risks, raising their betas and, hence, average returns in the following year. The third column in table 7 shows that this is not the case. The cross-sectional variation in fund average returns has nothing to do with market betas. Just as in the case of individual stock returns, we have to understand fund returns with multifactor models, if at all.

The last column of table 7 presents alphas (intercepts, the part of average return not explained by the model) from a model with four factors—the market, the Fama–French HML and SMB factors, and a momentum factor, PR1YR, that is long NYSE stocks that did well in the last year and short NYSE stocks that did poorly in the last year. In general, one should object to the inclusion of so many factors and such ad-hoc factors. However, this is a *performance attribution* rather than an *economic explanation* use of a multifactor model. We want to know whether fund performance, and persistence in fund performance in particular, is due to persistent stock-picking skill or to mechanical strategies that investors could just as easily follow on their own, without paying the management costs associated with investing through a fund. For this purpose, it does not matter whether the "factors" represent true, underlying sources of macroeconomic risks.

The alphas in the last column of table 7 are almost all about 1 percent to 2 percent per year negative. Thus, Carhart's model explains that the persistence in fund performance is due to persistence in the underlying stocks, not persistent stock-picking skill. These results support the old conclusion that actively managed funds underperform mechanical indexing strategies. There is some remaining puzzling persistence, but it is all in the large *negative* alphas of the bottom one-tenth to bottom one-thirtieth of performers, which lose money year after year. Carhart also shows that the persistence of fund performance is due to momentum in the underlying stocks, rather than momentum funds. If, by good luck, a fund happened to pick stocks that went up last year, the portfolio will continue to go up a bit this year.

In sum, the new research does nothing to dispel the disappointing view of active management. However, we discover that passively managed "style"

TABLE 7

Portfolios of mutual funds formed on previous year's return

Last year rank	Average return	CAPM alpha	4-factor alpha
	(---------- percent ----------)		
1/30	0.75	0.27	–0.11
1/10	0.68	0.22	–0.12
5/10	0.38	–0.05	–0.14
9/10	0.23	–0.21	–0.20
10/10	0.01	–0.45	–0.40
30/30	–0.25	–0.74	–0.64

Notes: Each year, mutual funds are sorted into portfolios based on the previous year's return. The rank column gives the rank of the selected portfolio. For example, 1/30 is the best performing portfolio when funds are divided into 30 categories. Average return gives the average monthly return in excess of the T-bill rate of this portfolio of funds for the following year. Four-factor alpha gives the average return less the predictions of a multifactor model that uses the market, the Fama–French HML and SMB portfolios, and portfolio PR1YR which is long NYSE stocks that did well in the last year and short NYSE stocks that did poorly in the last year.
Source: Carhart (1997).

portfolios can earn returns that are not explained by the CAPM.

Catastrophe insurance

A number of prominent funds have earned very good returns (and others, spectacular losses) by following strategies such as *convergence trades* and implicit *put options*. These strategies may also reflect high average returns as compensation for nonmarket dimensions of risk. They have not been examined at the same level of detail as the value and small-cap strategies, so I offer a possible interpretation rather than a documented one.

Convergence trades take strong positions in very similar securities that have small price differences. For example, a 29.5-year Treasury bond typically trades at a slightly higher yield (lower price) than a 30-year Treasury bond. (This was the most famous bet placed by LTCM. See Lewis, 1999.) A convergence trade puts a strong short position on the expensive security and a strong long position on the cheap security. This strategy is often mislabeled an "arbitrage." However, the securities are similar, not identical. The spread between 29.5- and 30-year Treasury bonds reflects the lower liquidity of the shorter maturity and the associated difficulty of selling it in a financial panic. It is possible for this spread to widen. Nonetheless, panics are rare, and the average returns in all the years when they do not happen may more than make up for the spectacular losses when they do.

Put options protect investors from large price declines. The *volatility smile* in put option prices reflects the surprisingly high prices of such options, compared with the small probability of large market collapses (even when one calibrates the probability directly, rather than using the log-normal distribution of the Black–Scholes formula). Writers of out-of-the-money puts collect a fee every month; in a rare market collapse they will pay out a huge sum, but if the probability of the collapse is small enough, the average returns may be quite good.

All of these strategies can be thought of as *catastrophe insurance* (Hsieh and Fung, 1999). Most of the time they earn a small premium. Once in a great while they lose a lot, and they lose a lot in times of financial catastrophe, when most investors are really anxious that the value of their investments not evaporate. Therefore, it is economically plausible that these strategies can earn positive average returns, even when we account for stock market risk via the CAPM and we correctly measure the small probabilities of large losses.

The difficulty in empirically estimating the true average return of such strategies, of course, is that rare events are rare. Many long samples will give a false sense of security because "the big one" that justifies the premium happened not to hit.

The value, yield curve, and foreign exchange strategies I survey above also exhibit features of catastrophe insurance. Value stocks may earn high returns because distressed stocks will all go bankrupt in a financial panic. Buying bonds of countries with high interest rates leaves one open to the small chance of a large devaluation, and such devaluation is especially likely to happen in a global financial panic. Similarly, buying long-term bonds in the depth of a recession when the yield curve is upward sloping may expose one to a small risk of a large inflation.

If these interpretations bear out, they also suggest that the premiums—the average returns from holding stocks sensitive to HML or from following the bond and foreign exchange strategies—may be overstated in the data. The markets have had an unusually good 50 years, and devastating financial panics have not happened.

Implications of the new facts

While the list of new facts appears long, similar patterns show up in every case. Prices reveal slow-moving market expectations of subsequent returns, because potential offsetting events seem sluggish or absent. The patterns suggest that investors can earn substantial average returns by taking on the risks of recession and financial stress. In addition, there is a small positive autocorrelation of high-frequency returns.

The effects are not completely new. We have known since the 1960s that high-frequency returns are slightly predictable, with R^2 of 0.01 to 0.1 in daily to monthly returns. These effects were dismissed because there didn't seem to be much one could do about them. A 51/49 bet is not very attractive, especially if there is any transaction cost. Also, the increased Sharpe ratio (mean excess return/standard deviation) from exploiting predictability is directly related to the forecast R^2, so a tiny R^2, even if exploitable, did not seem important. Now, we have a greater understanding of the potential importance of these effects and their economic interpretations.

For price effects, we now realize that the R^2 rises with horizon when the forecasting variables are slow-moving. Hence, a small R^2 at short horizons can mean a really substantial R^2 in the 30 percent to 50 percent range at longer horizons. Also, the nature of these effects suggests the kinds of additional sources of priced risk that theorists had anticipated for 20 years. For momentum effects, the ability to sort stocks and

funds into momentum-based portfolios means that very small predictability times portfolios with huge past returns gives important subsequent returns, though it is not totally clear that this amplification of the small predictability really does survive transaction costs.

Price-based forecasts

If expected returns rise, prices are driven down, since future dividends or other cash flows are discounted at a higher rate. A "low" price, then, can *reveal* a market expectation of a high expected or required return.[10]

Most of our results come from this effect. Low price/dividend, price/earnings, or price/book values signal times when the market as a whole will have high average returns. Low market value (price times shares) relative to book value signals securities or portfolios that earn high average returns. The "small-firm" effect derives from low prices—other measures of size such as number of employees or book value alone have no predictive power for returns (Berk, 1997). The "five-year reversal" effect derives from the fact that five years of poor returns lead to a low price. A high long-term bond yield means that the price of long-term bonds is "low," and this seems to signal a time of good long-term bond returns. A high foreign interest rate means a low price on foreign bonds, and this seems to indicate good returns on the foreign bonds.

The most natural interpretation of all these effects is that the expected or required return—the risk premium—on individual securities as well as the market as a whole varies slowly over time. Thus we can track market expectations of returns by watching price/dividend, price/earnings, or book/market ratios.

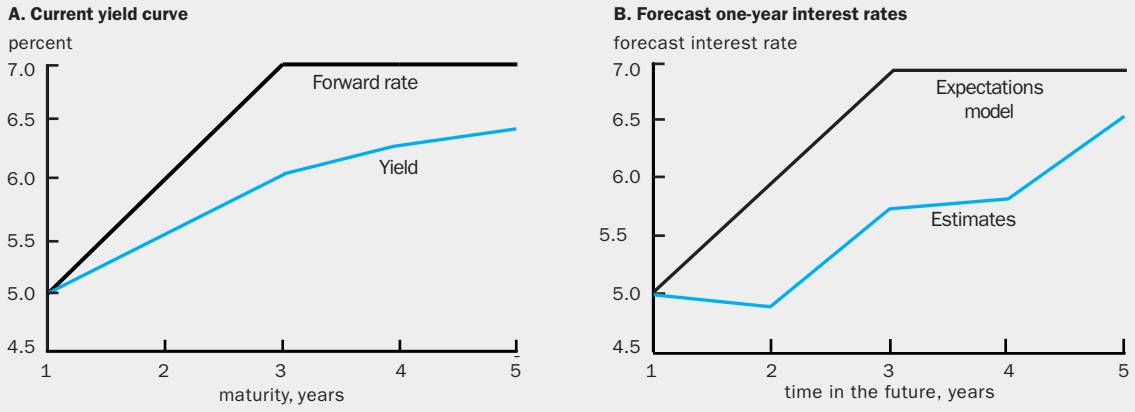
Absent offsetting events

In each case, an apparent difference in yield should give rise to an offsetting movement, but does not seem to do so. Something *should* be predictable so that returns are *not* predictable, and it is not. Figure 8 provides a picture of the results in table 5. Suppose that the yield curve is upward sloping as in panel A. What does this mean? If the expectations model were true, the forward rates plotted against maturity would translate one for one to the forecast of future spot rates in panel B, as plotted in the black line marked "Expectations model." A high long-term bond yield relative to short-term bond yields should not mean a higher expected long-term bond return. Subsequent short rates should rise, cutting off the one-period advantage of long-term bonds and raising the multi-year advantage of short-term bonds.

In figure 8, panel b, the colored line marked "Estimates" shows the actual forecast of future spot interest rates from the results in table 5. The essence of the phenomenon is *sluggish adjustment* of the short rates. The short rates do eventually rise to meet the forward rate forecasts, but not as quickly as the forward rates predict they should. Short-term yields *should* be forecastable so that returns are *not* forecastable. In fact, yields are almost unforecastable, so, mechanically, bond returns are. The roughly 1.0 coefficients in panel B of table 5 mean that a 1 percentage point increase in the forward rate translates into a 1 percentage point increase in expected return. It seems that old fallacy of confusing bond *yields* with their expected *returns* for the first year contains a grain of truth.

FIGURE 8

Yield curve and forecast one-year interest rates



Notes: Assuming that the current yield curve is as shown in panel A, the black line in panel B gives the forecast from the expectations hypothesis, in which case forward rates today are the forecast of future spot rates. The colored line in panel B gives the actual forecast of future spot rates, constructed from the estimates in table 5.

In the same way, a high dividend yield on a stock or portfolio should mean that dividends grow more slowly over time, or, for individual stocks, that the firm has taken on more market risk and will have a higher market beta. These tendencies seem to be completely absent. Dividend/price ratios do not seem to forecast dividend growth and, hence, (mechanically) they forecast returns. The one-year coefficient in table 1 is very close to 1.00, meaning that a 1 percentage point increase in the dividend yield translates into a 1 percentage point increase in return. It seems that the old fallacy of confusing increased dividend yield with increased total return does contain a grain of truth.

A high foreign interest rate relative to domestic interest rates should not mean a higher expected return. We should see, on average, an offsetting depreciation. But here, the coefficients are even larger than 1.0. An interest rate differential seems to predict a further *appreciation.* It seems that the old fallacy of confusing interest rate differentials across countries with expected returns, forgetting about depreciation, also contains a grain of truth.

Economic interpretation

The price-based predictability patterns suggest a premium for holding risks related to recession and economy-wide financial distress. Stock and bond predictability are linked: The term spread (forward-spot, or long yield–short yield) forecasts stock returns as well as bond returns (Fama and French, 1989). Furthermore, the term spread is one of the best variables for forecasting business cycles. It rises steeply at the bottom of recessions and is inverted at the top of a boom. Return forecasts are high at the bottom of a business cycle and low at the top of a boom. Value and small-cap stocks are typically distressed. Empirically successful economic models of the recession and distress premiums are still in their infancy (Campbell and Cochrane, 1999, is a start), but the story is at least plausible and the effects have been expected by theorists for a generation.

To make this point come to life, think concretely about what you have to do to take advantage of the predictability strategies. You have to buy stocks or long-term bonds at the bottom, when stock prices are low after a long and depressing bear market, in the bottom of a recession or the peak of a financial panic. This is a time when few people have the guts or the wallet to buy risky stocks or risky long-term bonds. Looking across stocks rather than over time, you have to invest in value or small-cap companies, with years of poor past returns, poor sales, or on the edge of bankruptcy. You have to buy stocks that everyone else thinks are dogs. Then, you have to sell stocks and long-term bonds in good times, when stock prices are high relative to dividends, earnings, and other multiples and the yield curve is flat or inverted so that long-term bond prices are high. You have to sell the popular growth stocks, with good past returns, good sales, and earnings growth.

You have to sell now, and the stocks that you should sell are the blue-chips that everyone else seems to be buying. In fact, the market timing strategies said to sell long ago; if you did so, you would have missed much of the runup in the Dow past the 6,000 point. Value stocks too have missed most of the recent market runup. However, this shouldn't worry you—a strategy that holds risks uncorrelated with the market must underperform the market close to half of the time.

If this feels uncomfortable, what you're feeling is risk. If you're uncomfortable watching the market pass you by, perhaps you *don't* really only care about long-run mean and variance; you also care about doing well when the market is doing well. If you want to stay fully invested in stocks, perhaps you too feel the time-varying aversion to or exposure to risk that drives the average investor to stay fully invested despite low prospective returns.

This line of explanation for the foreign exchange puzzle is still a bit farther off (see Engel, 1996, for a survey; Atkeson, Alvarez, and Kehoe, 1999, offer a recent stab at an explanation). The strategy leads investors to invest in countries with high interest rates. High interest rates are often a sign of monetary instability or other economic trouble, and thus may mean that the investments are more exposed to the risks of global financial stress or a global recession than are investments in the bonds of countries with low interest rates, which are typically enjoying better times.

Return correlation

Momentum and persistent fund performance explained by a momentum factor are different from the price-based predictability results. In both cases, the underlying phenomenon is a small predictability of high-frequency returns. The price-based predictability strategies make this predictability important by showing that, with a slow-moving forecasting variable, the R^2 builds over horizon. Momentum, however, is based on a fast-moving forecast variable—the previous year's return. Therefore, the R^2 declines rather than building with horizon. Momentum makes the small predictability of high-frequency returns significant in a different way, by forming portfolios of extreme winners and losers. The large volatility of returns means that the extreme portfolios will have extreme past returns, so only a small continuation of past returns gives a large current return.

It would be appealing to understand momentum as a reflection of slowly time-varying average expected returns or risk premiums, like the price-based predictability strategies. If a stock's average return rises for a while, that should make returns higher both today and tomorrow. Thus, a portfolio of past winners will contain more than its share of stocks that performed well because their average returns were higher, along with stocks that performed well due to luck. The average return of such a portfolio should be higher tomorrow as well.

Unfortunately, this story has to posit a substantially different view of the underlying process for varying expected returns than is needed to explain everything else. The trouble is that a surprise increase in expected returns means that prices will fall, since dividends are now discounted at a greater rate. This is the phenomenon we have relied on to explain why *low* price/dividend, price/earnings, book/market, value, and size forecast *higher* subsequent returns. Therefore, *positive* correlation of *expected* returns typically yields a *negative* correlation of *realized* returns. To get a positive correlation of realized returns out of slow expected return variation, you have to imagine that an increase in average returns today is either highly correlated with a decrease in expected future dividend growth or with a decrease in expected returns in the distant future (an impulse response that starts positive but is negative at long horizons). Campbell, Lo, and MacKinlay (1997) provide a quantitative exposition of these effects.

Furthermore, momentum returns have not yet been linked to business cycles or financial distress in even the informal way that I suggested for price-based strategies. Thus, momentum still lacks a plausible economic interpretation. To me, this adds weight to the view that it isn't there, it isn't exploitable, or it represents a small illiquidity (tax-loss selling of small illiquid stocks) that will be quickly remedied once a few traders understand it.

Remaining doubts

The size of all these effects is still somewhat in question. It is always hard to measure average returns of risky strategies. The standard formula $\sigma / \sqrt{T}$ for the standard error of the mean, together with the high volatility σ of any strategy, means that one needs 25 years of data to even start to measure average returns. With $\sigma = 16$ percent (typical of the index), even $T = 25$ years means that one standard error is $16/5 \cong 3$ percent per year, and a two-standard error confidence interval runs plus or minus 6 percentage points. This is not much smaller than the average returns we are trying to measure. In addition, all of these facts are highly influenced by the small probability of rare events, which makes measuring average returns even harder.

Finally, viewed the right way, we have very few data points with which to evaluate predictability. The term premium and interest rate differentials only change sign with the business cycle, and the dividend/price ratio only crosses its mean once every generation. The history of interest rates and inflation in the U.S. is dominated by the increase, through two recessions, to a peak in 1980 and then a slow decline after that.

Many of the anomalous risk premiums seem to be declining over time. Figure 6 shows the decline in the HML and SMB premiums, and the same may be true of the predictability effects. The last three years of high market returns have cut the estimated return predictability from the dividend/price ratio in *half*. This fact suggests that at least some of the premium the new strategies yielded in the past was due to the fact that they were simply overlooked.

Was it really clear to average investors in 1947 or 1963 (the beginning of the data samples) that stocks would earn 9 percent over bonds, and that the strategy of buying distressed small stocks would double even that return for the same level of risk? Would average investors have changed their portfolios with this knowledge? Or would they have stayed pat, explaining that these returns are earned as a reward for risk that they were not willing to take? Was it clear that buying stocks at the bottom in the mid-1970s would yield so much more than even that high average return? If we interpret the premiums measured in sample as true risk premiums, the answer must be yes. If the answer is no, then at least some part of the premium was luck and will disappear in the future.

Since the premiums are hard to measure, one is tempted to put less emphasis on them. However, they are crucial to our interpretation of the facts. The CAPM is perfectly consistent with the fact that there are additional sources of *common* variation. For example, it was long understood that stocks in the same industry move together; the fact that value or small stocks also move together need not cause a ripple. The surprise is that investors seem to earn an average return premium for holding these additional sources of common movement, whereas the CAPM predicts that (given beta) they should have no effect on a portfolio's average returns.

The behavior of funds also suggests the "overlooked strategy" interpretation. As explained earlier, fund returns still cluster around the market line. It turns out that very few fund returns actually followed the value or other return-enhancing strategies. However, the number of small, value, and related funds—funds

that actually do follow the strategies—has increased dramatically in recent years. It might be possible to explain this in a way consistent with the idea that investors knew the premiums were there all along, but such an argument is obviously strained.

Conclusion

In sum, it now seems that investors can earn a substantial premium for holding dimensions of risk unrelated to market movements, such as recession-related or distress-related risk. Investors earn these premiums by following strategies, such as value and growth, market-timing possibilities generated by return predictability, dynamic bond and foreign exchange strategies, and maybe even a bit of momentum. The exact size of the premiums and the economic nature of the underlying risks is still a bit open to question, but researchers are unlikely to go back to the simple view that returns are independent over time and that the CAPM describes the cross section.

The next question is, What should investors do with this information? The article, "Portfolio advice for a multifactor world," also in this issue, addresses that question.

NOTES

[1]The market also tends to go down in recessions; however recessions can be unusually severe or mild for a given level of market return. What counts here is the severity of the recession for a given market return. Technically, we are considering betas in a multiple regression that includes both the market return and a measure of recessions. See box 1.

[2]I thank Gene Fama for providing me with these data.

[3]The rest of the paragraph is my interpretation, not Fama and French's. They focus on the firm's financial distress, while I focus on the systematic distress, since idiosyncratic distress cannot deliver a risk price.

[4]Fama and French do not provide direct measures of standard deviations for these portfolios. One can infer, however, from the betas, R^2 values, and standard deviation of the market and factor portfolios that the standard deviations are roughly one to two times that of the market return, so Sharpe ratios of these strategies are comparable to that of the market return in sample.

[5]We are looking for

$$E(r|r \geq x) = \frac{\int_x^\infty r f(r)dr}{\int_x^\infty f(r)dr},$$

where x is defined as the top one-tenth cutoff,

$$\int_x^\infty f(r)dr = \frac{1}{10}.$$

With a normal distribution, $x = 1.2816\sigma$ and $E(r|r \geq x) = 1.755\sigma$.

[6]The index autocorrelations suffer from some upward bias since some stocks do not trade every day. Individual stock autocorrelations are generally smaller, but are enough to account for the momentum effect.

[7]Panel B is really not independent evidence, since the coefficients in panels A and B of table 5 are mechanically linked. For example, 1.14 + (–0.14) = 1.0, and this holds as an accounting identity. Fama and Bliss (1987) call them "complementary regressions."

[8]As with bonds, the expectations hypothesis is slightly different from pure risk neutrality since the expectation of the log is not the log of the expectation. Again, the size of the phenomena we study swamps this distinction.

[9]The data are actually the spread between the forward exchange rate and the spot exchange rate, but this quantity must equal the interest rate differential in order to preclude arbitrage.

[10]This effect is initially counterintuitive. One might suppose that a higher average return would attract investors, raising prices. But the higher prices, for a given dividend stream, must reduce subsequent average returns. High average returns persist, in equilibrium, when investors fear the increased risks of an asset and try to sell, lowering prices.

REFERENCES

Atkeson, Andrew, Fernando Alvarez, and Patrick Kehoe, 1999, "Volatile exchange rates and the forward premium anomaly: A segmented asset market view," University of Chicago, working paper.

Banz, R. W., 1981, "The relationship between return and market value of common stocks," *Journal of Financial Economics,* Vol. 9, No. 1, pp. 3–18.

Bekaert, Geert, and Robert J. Hodrick, 1992, "Characterizing predictable components in excess returns on equity and foreign exchange markets," *Journal of Finance*, Vol. 47, No. 2, June, pp. 467–509.

Berk, Jonathan, 1997, "Does size really matter?," *Financial Analysts Journal*, Vol. 53, September/October, pp. 12–18.

Campbell, John Y., 1996, "Understanding risk and return," *Journal of Political Economy*, Vol. 104, No. 2, April, pp. 298–345.

Campbell, John Y., and John H. Cochrane, 1999, "By force of habit: A consumption-based explanation of aggregate stock market behavior," *Journal of Political Economy*, Vol. 107, No. 2, April, pp. 205–251.

Campbell, John Y., Andrew W. Lo, and A. Craig MacKinlay, 1997, *The Econometrics of Financial Markets,* Princeton, NJ: Princeton University Press.

Carhart, Mark M., 1997, "On persistence in mutual fund performance," *Journal of Finance*, Vol. 52, No. 1, March, pp. 57–82.

Chen, Nai-Fu, Richard Roll, and Stephen A. Ross, 1986, "Economic forces and the stock market," *Journal of Business*, Vol. 59, No. 3, July, pp. 383–403.

Cochrane, John H., 1997, "Where is the market going? Uncertain facts and novel theories," *Economic Perspectives*, Federal Reserve Bank of Chicago, Vol. 21, No. 6, November/December, pp. 3–37.

_____, 1996, "A cross-sectional test of an investment-based asset pricing model," *Journal of Political Economy*, Vol. 104, No. 3, June, pp. 572–621.

_____, 1991, "Volatility tests and efficient markets: Review essay," *Journal of Monetary Economics*; Vol. 27, No. 3, June, pp. 463-485.

Daniel, Kent, David Hirshleifer, and Avanidhar Subrahmanyam, 1998, "Investor psychology and security market under- and overreactions," *Journal of Finance*, Vol. 3, No. 6, December, pp. 1839–1885.

DeBondt, Werner F. M., and Richard H. Thaler, 1985, "Does the stock market overreact?," *Journal of Finance*, Vol. 40, No. 3, pp. 793–805.

Engel, Charles, 1996, "The forward discount anomaly and the risk premium: A survey of recent evidence," *Journal of Empirical Finance*, Vol. 3, pp. 123–192.

Fama, Eugene F. 1991, "Efficient markets II," *Journal of Finance*, Vol. 46, No. 5, December, pp. 1575–1617.

_____, 1984, "Forward and spot exchange rates," *Journal of Monetary Economics*, Vol. 14, No. 3, November, pp. 319–338.

_____, 1970, "Efficient capital markets: A review of theory and empirical work," *Journal of Finance*, Vol. 25, No. 2, May, pp. 383–417.

_____, 1965, "The behavior of stock market prices," *Journal of Business*, Vol. 38, No. 1, pp. 34–105.

Fama, Eugene F., and Robert R. Bliss, 1987, "The information in long-maturity forward rates," *American Economic Review*, Vol. 77, No. 4, September, pp. 680–692.

Fama, Eugene F., and Kenneth R. French, 1997, "Industry costs of equity," *Journal of Financial Economics*, Vol. 43, No. 2, February, pp. 153–193.

_____, 1996, "Multifactor explanations of asset-pricing anomalies," *Journal of Finance*, Vol. 51, No. 1, March, pp, 55–84.

_____, 1995, "Size and book-to-market factors in earnings and returns," *Journal of Finance*, Vol. 50, No.1, March, pp. 131–155.

_____, 1993, "Common risk factors in the returns on stocks and bonds," *Journal of Financial Economics*, Vol. 33, No. 1, February, pp. 3–56.

_____, 1989, "Business conditions and expected returns on stocks and bonds," *Journal of Financial Economics*, Vol. 25, No. 1, November, pp. 23–49.

Heaton, John, and Deborah Lucas, 1997, "Portfolio choice and asset prices: The importance of entrepreneurial risk," Northwestern University, manuscript.

Hendricks, Darryll, Jayendu Patel, and Richard Zeckhauser, 1993, "Hot hands in mutual funds: Short-term persistence of performance," *Journal of Finance,* Vol. 48, No. 1, March, pp. 93–130.

Hodrick, Robert, 2000, *International Financial Management,* Englewood Cliffs, NJ: Prentice-Hall, forthcoming.

Hsieh, David, and William Fung, 1999, "Hedge fund risk management," Duke University, working paper.

Jagannathan, Ravi, and Zhenyu Wang, 1996, "The conditional CAPM and the cross-section of expected returns," *Journal of Finance*, Vol. 51, No. 1, March, pp. 3–53.

Jegadeesh, Narasimham, and Sheridan Titman, 1993, "Returns to buying winners and selling losers: Implications for stock market efficiency," *Journal of Finance*, Vol. 48, No. 1, March, pp. 65–91.

Jensen, Michael C., 1969, "The pricing of capital assets and evaluation of investment portfolios," *Journal of Business*, Vol. 42, No. 2, April, pp. 167–247.

Lakonishok, Josef, Andrei Shleifer, and Robert W. Vishny, 1992, "The structure and performance of the money management industry," *Brookings Papers on Economic Activity: Microeconomics 1992*, Washington, DC, pp. 339–391

LeRoy, Stephen F., and Richard D. Porter, 1981, "The present-value relation: Tests based on implied variance bounds," *Econometrica*, Vol. 49, No. 3, May, pp. 555–574.

Lewis, Karen, K., 1995, "Puzzles in international financial markets," in *Handbook of International Economics,* Vol. 3, G. Grossman and K. Rogoff (eds.), Amsterdam, New York, and Oxford: Elsevier Science B.V, pp. 1913–1971.

Lewis, Michael, 1999, "How the eggheads cracked," New *York Times Magazine*, January 24, pp. 24–42.

Liew, Jimmy, and Maria Vassalou, 1999, "Can book-to-market, size and momentum be risk factors that predict economic growth?," Columbia University, working paper.

MacKinlay, A. Craig, 1995, "Multifactor models do not explain deviations from the CAPM," *Journal of Financial Economics*, Vol. 38, No. 1, pp. 3–28.

Malkiel, Burton, 1990, *A Random Walk Down Wall Street*, New York: Norton.

Markowitz, H., 1952, "Portfolio selection," *Journal of Finance*, Vol. 7, No. 1, March, pp. 77–99.

Meredith, Guy, and Menzie D. Chinn, 1998, "Long-horizon uncovered interest rate parity," National Bureau of Economic Research, working paper, No. 6797.

Merton, Robert C., 1973, "An intertemporal capital asset pricing model," *Econometrica*, Vol. 41, No. 5, September, pp. 867–887.

_____, 1971, "Optimum consumption and portfolio rules in a continuous time model," *Journal of Economic Theory*, Vol. 3, No. 4, pp. 373–413.

_____, 1969, "Lifetime portfolio selection under uncertainty: The continuous time case," *Review of Economics and Statistics*, Vol. 51, No. 3, August, pp. 247–257.

Moskowitz, Tobias, and Mark Grinblatt, 1999, "Tax loss selling and return autocorrelation: New evidence," University of Chicago, working paper.

_____, 1998, "Do industries explain momentum?," University of Chicago, CRSP working paper, No. 480.

New York Times Company, 1999, "Mutual funds report: What's killing the value managers?," *New York Times*, April 4, Section 3, p. 29.

Reyfman, Alexander, 1997, "Labor market risk and expected asset returns," University of Chicago, Ph.D. thesis.

Ross, S. A., 1976, "The arbitrage theory of capital asset pricing," *Journal of Economic Theory*, Vol. 13, No. 3, December, pp. 341–360.

Samuelson, Paul A., 1969, "Lifetime portfolio selection by dynamic stochastic programming," *Review of Economics and Statistics*, Vol. 51, No. 3, August, pp. 239–246.

Sargent, Thomas J., 1993, *Bounded Rationality in Macroeconomics*, Oxford: Oxford University Press.

Shiller, Robert J., 1981, "Do prices move too much to be justified by subsequent changes in dividends?," *American Economic Review*, Vol. 71, No. 3, June, pp. 421–436.

Portfolio advice for a multifactor world

John H. Cochrane

Introduction and summary

A companion article in this issue, "New facts in finance," summarizes the revolution in how financial economists view the world. Briefly, there are strategies that result in high average returns without large *betas*, that is, with no strong tendency for the strategy's returns to move up and down with the market as a whole. Multifactor models have supplanted the capital asset pricing model (CAPM) in describing these phenomena. Stock and bond returns, once thought to be independent over time, turn out to be predictable at long horizons. All of these phenomena seem to reflect a premium for holding macroeconomic risks associated with the business cycle and for holding assets that do poorly in times of financial distress. They also all reflect the information in prices—high prices lead to low returns and low prices lead to high returns.

The world of investment opportunities has also changed. Where once an investor faced a fairly straightforward choice between managed mutual funds, index funds, and relatively expensive trading on his own account, now he must choose among a bewildering variety of fund *styles* (such as value, growth, small cap, balanced, income, global, emerging market, and convergence), as well as more complex claims of active fund managers with customized styles and strategies, and electronic trading via the Internet. (Msn.com's latest advertisement suggests that one should sign up in order to "check the hour's hottest stocks." Does a beleaguered investor really have to do that to earn a reasonable return?) The advertisements of investment advisory services make it seem important to tailor an investment portfolio from this bewildering set of choices to the particular circumstances, goals, and desires of each investor.

What should an investor do? An important current of academic research investigates how portfolio theory should adapt to our new view of the financial world. In this article, I summarize this research and I distill its advice for investors. In particular, which of the bewildering new investment styles seem most promising? Should you attempt to time stock, bond, or foreign exchange markets, and if so how much? To what extent and how should an investment portfolio be tailored to your specific circumstances? Finally, what can we say about the future investment environment? What kind of products will be attractive to investors in the future, and how should public policy react to these financial innovations?

I start by reviewing the traditional academic portfolio advice, which follows from the traditional view that the CAPM is roughly correct and that returns are not predictable over time. In that view, all investors (who do not have special information) should split their money between risk-free bonds and a broad-based passively managed index fund that approximates the "market portfolio." More risk-tolerant investors put more money into the stock fund, more risk averse investors put more money into the bond fund, and that is it.

The new academic portfolio advice reacts to the new facts. An investor should hold, in addition to the market portfolio and risk-free bonds, a number of passively managed "style" funds that capture the broad (nondiversifiable) risks common to large numbers

John H. Cochrane is the Sigmund E. Edelstone Professor of Finance in the Graduate School of Business at the University of Chicago, a consultant to the Federal Reserve Bank of Chicago, and a research associate at the National Bureau of Economic Research (NBER). *The author thanks Andrea Eisfeldt for research assistance and David Marshall, John Campbell, and Robert Shiller for comments. The author's research is supported by the Graduate School of Business and by a grant from the National Science Foundation, administered by the NBER.*

of investors. In addition to the overall level of risk aversion, his exposure to or aversion to the various additional risk factors matters as well. For example, an investor who owns a small steel company should shade his investments away from a steel industry portfolio, or cyclical stocks in general; a wealthy investor with no other business or labor income can afford to take on the "value" and other stocks that seem to offer a premium in return for potentially poor performance in times of financial distress. The stock market is a way of transferring risks; those exposed to risks can hedge them by proper investments, and those who are not exposed to risks can earn a premium by taking on risks that others do not wish to shoulder.

Since returns are somewhat predictable, investors can enhance their average returns by moving their assets around among broad categories of investments. However, the market timing signals are slow-moving, and I show that the uncertainty about the nature and strength of market timing effects dramatically reduces the optimal amount by which investors can profit from them.

I emphasize a cautionary fact: *The average investor must hold the market.* You should only vary from a passive market index if you are different from everyone else. It cannot be the case that every investor should tilt his portfolio toward "value" or other high-yield strategies. If everybody did it, the phenomenon would disappear. Thus, if such strategies will persist at all, it must be the case that for every investor who should take advantage of them, there is another investor who should take an unusually *low* position, sacrificing the good average returns for a reduction in risk. It cannot be the case that every investor should "market time," buying when prices are low and selling when prices are high. If everyone did it, that phenomenon would also disappear. The phenomenon can only persist if, for every investor who should enhance returns by such market-timing, there is another investor who is so exposed to or averse to the time-varying risks that cause return predictability, that he *should* "buy high and sell low," again earning a lower average return in exchange for avoiding risks.

We have only scratched the surface of asset markets' usefulness for sharing risks. As often in economics, what appears from the outside to be greedy behavior is in fact socially useful.

The traditional view

The new portfolio theory really extends rather than overturns the traditional academic portfolio theory. Thus, it's useful to start by reminding ourselves what the traditional portfolio theory is and why. The traditional academic portfolio theory, starting from Markowitz (1952) and expounded in every finance textbook, remains one of the most useful and enduring bits of economics developed in the last 50 years.

Two-fund theorem

The traditional advice is to split your investments between a money-market fund and a broad-based, passively managed stock fund. That fund should concentrate on minimizing fees and transaction costs, period. It should avoid the temptation to actively manage its portfolio, trying to chase the latest hot stock or trying to foresee market movements. An index fund or other approximation to the *market portfolio* that passively holds a bit of every stock is ideal.

Figure 1 summarizes the analysis behind this advice. The straight line gives the *mean-variance frontier*—the portfolios that give the highest mean return for every level of volatility. Every investor should pick a portfolio on the mean-variance frontier. The upward-curved lines are *indifference curves* that represent investors' preference for more mean return and less volatility. The indifference curve to the lower left represents a risk-averse investor, who chooses a portfolio with less mean return but also less volatility; the indifference curve to the upper right represents a more risk-tolerant investor who chooses a portfolio with more mean return but also higher risk.

This seems like a lot of person-specific portfolio formation. However, every portfolio on the mean-variance frontier can be formed as a combination of a risk-free money-market fund and the *market portfolio* of all risky assets. Therefore, every investor need only hold different proportions of these *two funds*.

Bad portfolio advice

The portfolio advice is not so remarkable for what it does say, which given the setup is fairly straightforward, as for what it does not say. Compared with common sense and much industry practice, it is radical advice.

One might have thought that investors willing to take on a little more risk in exchange for the promise of better returns should weight their portfolios to riskier stocks, or to value, growth, small-cap, or other riskier fund styles. Conversely, one might have thought that investors who are willing to forego some return for more safety should weight their portfolios to safer stocks, or to blue-chip, income, or other safer fund styles. Certainly, some professional advice in deciding which style is suited for an investor's risk tolerance, if not a portfolio professionally tailored to each investor's circumstances, seems only sensible and prudent. The advertisements that promise "we build the portfolio that's right for *you*" cater to this natural and sensible-sounding idea.

Figure 1 proves that nothing of the sort is true. All stock portfolios lie on or inside the curved risky asset frontier. Hence, an investor who wants more return and is willing to take more risk than the market portfolio will do better by borrowing to invest in the market—including the large-cap, income, and otherwise safe stocks—than by holding a portfolio of riskier stocks. An investor who wants something less risky than the market portfolio will do better by splitting his investment between the market and a money-market fund than by holding only safe stocks, even though his stock portfolio will then contain some of the small-cap, value, or otherwise risky stocks. Everyone holds the same market portfolio; the only decision is how much of it to hold.

The two-fund theorem in principle still allows for a good deal of customized portfolio formation and active management if investors or managers have different *information* or *beliefs*. For example, if an investor knows that small-cap stocks are ready for a rebound, then the optimal (or tangency) portfolio that reflects this knowledge will be more heavily weighted toward small-cap stocks than the market portfolio held by the average investor. All the analysis of figure 1 holds, but this specially constructed *tangency portfolio* goes in the place indicated by the market portfolio in the figure. However, the empirical success of market efficiency, and the poor performance of professional managers relative to passive indexation, strongly suggests that these attempts will not pay off for most investors. For this reason, the standard advice is to hold passively managed funds that concentrate on minimizing transaction costs and fees, rather than a carefully constructed tangency portfolio that reflects an investor's or manager's special insights. However, a quantitative portfolio management industry tries hard to mix information or beliefs about the behavior of different securities with the theory of figure 1 (for example, see Black and Litterman, 1991).

The two-fund theorem leaves open the possibility that the investor's *horizon* matters as well as his risk aversion. What could be more natural than the often repeated advice that a long-term investor can afford to ride out all the market's short-term volatility, while a short-term investor should avoid stocks because he may have to sell at the bottom rather than wait for the inevitable recovery after a price drop? The fallacy lies in the "inevitable" recovery. If returns are close to independent over time (like a coin flip), and prices are close to a random walk, a price drop makes it no more likely that prices will rise more in the future. This view implies that stocks are *not* safer in the long run, and the stock/bond allocation should be independent of investment horizon.

This proposition can be shown to be precisely true in several popular mathematical models of the portfolio decision. If returns are independent over time, then the mean and variance of continuously compounded returns rises in proportion to the horizon: The mean and variance of ten-year returns are ten times those of one-year returns, so the ratio of mean to variance is the same at all horizons. More elegantly, Merton (1969) and Samuelson (1969) show that an investor with a constant relative risk aversion utility who can continually rebalance his portfolio between stocks and bonds will always choose the same stock/bond proportion regardless of investment horizon, when returns are independent over time.

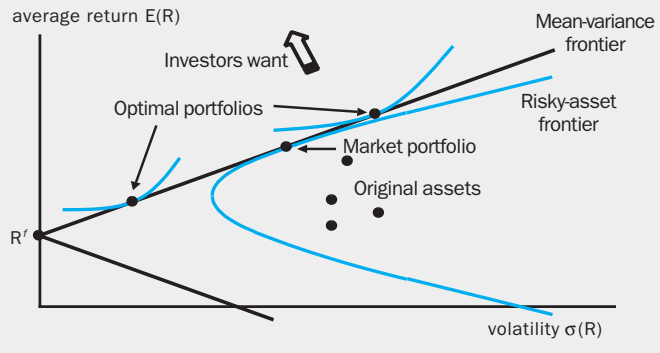
Taking the advice

This advice has had a sizable impact on portfolio practice. Before this advice was widely popularized in the early 1970s, the proposition that professional active management and stock selection could outperform blindly holding an index seemed self-evident, and passively managed index funds were practically unknown. They have exploded in size since then. The remaining actively managed funds clearly feel the need to defend active management in the face of the advice to hold passive index funds and the fact that active managers selected on any ex-ante basis underperform indexes ex-post.

The one input to the optimal portfolio advice is risk tolerance, and many providers of investment services have started thinking about how to measure risk tolerance using a series of questionnaires. This is the trickiest part of the conventional advice, in part since conventional

FIGURE 1

Mean-variance frontier, optimal portfolios and two-fund theorem



measures of risk tolerance often seem quite out of whack with risk aversion displayed in asset markets. (This is the *equity premium puzzle*; see Cochrane, 1997, for a review.) However, the basic question is whether one is more risk tolerant or less risk tolerant than the average investor. This question is fairly easy to conceptualize and can lead to a solid qualitative, if not quantitative, answer.

One might object to the logical inconsistency of providing portfolio advice based on a view of the world in which everyone is already following such advice. (This logic is what allowed me to identify the mean-variance frontier with the market portfolio.) However, this logic is only wrong if other investors are *systematically* wrong. If some investors hold too much of a certain stock, but others hold too little of it, market valuations are unaffected and the advice to hold the market portfolio is still valid.

New portfolio theory

Multiple factors: An N-fund theorem

Figure 2 shows how the simple two-fund theorem of figure 1 changes if there are multiple sources of priced risk. This section is a graphical version of Fama's (1996) analysis. Much of the theory comes from Merton (1969, 1971, 1973).

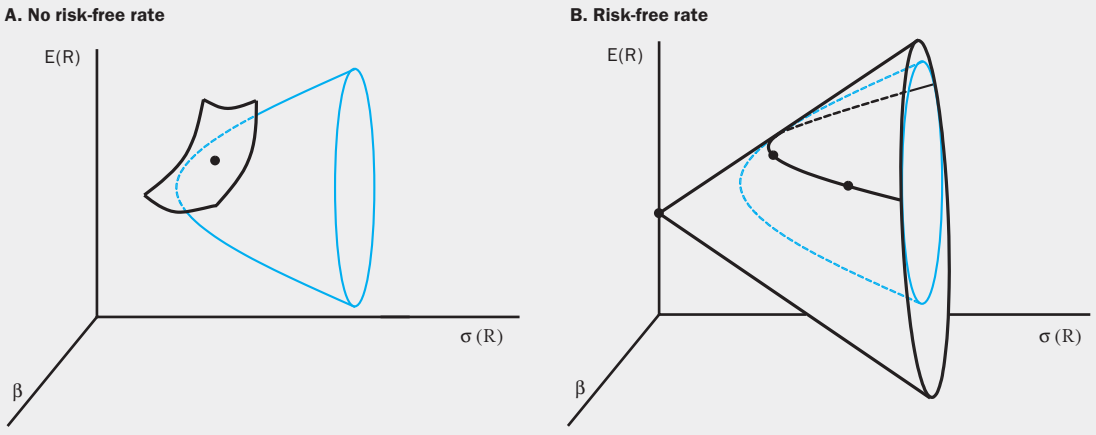
To keep the figure simple I consider just one additional factor. For concreteness, think of an additional recession factor. Now, investors care about three attributes of their portfolios: 1) They want higher average returns; 2) they want lower standard deviations or overall risk; and 3) they are willing to accept a portfolio with a little lower mean return or a little higher standard deviation of return if the portfolio does not do poorly in recessions. In the context of figure 2, this means that investors are happier with portfolios that are higher up (more mean), more to the left (less standard deviation), and farther out (lower recession sensitivity). The indifference *curves* of figure 1 become indifference *surfaces*. Panel A of figure 2 shows one such surface curving upwards.

As with figure 1, we must next think about what is available. We can now calculate a frontier of portfolios based on their mean, variance, and recession sensitivity. This frontier is the *multifactor efficient* frontier. A typical investor then picks a point as shown in panel A of figure 2, which gives him the best possible portfolio—trading off mean, variance, and recession sensitivity—that is available. Investors want to hold multifactor efficient, rather than mean-variance efficient, portfolios. As the mean-variance frontier of figure 1 is a hyperbola, this frontier is a revolution of a hyperbola. The appendix summarizes the mathematics behind this figure.

Panel B of figure 2 adds a risk-free rate. As the mean-variance frontier of figure 1 was the minimal V shape emanating from the risk-free rate (R^f) that includes the hyperbolic risky frontier, now the multifactor efficient frontier is the minimum *cone* that includes the hyperbolic risky multifactor efficient frontier, as shown.

FIGURE 2

Portfolio theory in a multifactor world



Notes: Panel A shows an indifference surface and optimal portfolio in the case with no risk-free rate. The dot marks the optimal portfolio where the indifference surface touches the multifactor efficient frontier. Panel B shows the set of multifactor efficient portfolios with a risk-free rate. The two cone-shaped surfaces intersect on the black line with two dots. The two dots are the market portfolio and an additional multifactor-efficient portfolio; all multifactor-efficient portfolios on the outer cone can be reached by combinations of the risk-free rate, the market, and the extra multifactor-efficient portfolio.

As every point of the mean-variance frontier of figure 1 can be reached by some combination of two funds—a risk-free rate and the market portfolio—now every point on the multifactor efficient frontier can be reached by some combination of *three* multifactor efficient funds. The most convenient set of portfolios is the risk-free rate (money-market security), the market portfolio (the risky portfolio held by the average investor), and one additional multifactor efficient portfolio on the tangency region as shown in panel B of figure 2. It is convenient to take this third portfolio to be a zero-cost, zero-beta portfolio, so that it isolates the extra dimension of risk.

Investors now may differ in their desire or ability to take on recession-related risk as well as in their tolerance for overall risk. Thus, some will want portfolios that are farther in and out, while others will want portfolios that are farther to the left and right. They can achieve these varied portfolios by different weights in the *three* multifactor efficient portfolios, or *three funds*.

Implications for mean-variance investors

The mean-variance frontier still exists—it is the projection of the cone shown in figure 2 on the mean-variance plane. As the figure shows, the average investor is willing to give up some mean or accept more variance in order to reduce the recession-sensitivity of his portfolio. The average investor must hold the market portfolio, so *the market return is no longer on the mean-variance frontier*.

Suppose, however, that *you* are concerned only with mean and variance—you are not exposed to the recession risk, or the risks associated with any other factor, and you only want to get the best possible mean return for given standard deviation. If so, you still want to solve the mean-variance problem of figure 1, and you still want a mean-variance efficient portfolio. The important implication of a multifactor world is that you, the mean-variance investor, should no longer hold the *market* portfolio.

You can still achieve a mean-variance efficient portfolio just as in figure 1 by a combination of a money market fund and a single *tangency portfolio*, lying on the upper portion of the curved risky-asset frontier. The tangency portfolio now takes stronger positions than the market portfolio in factors such as value or recession-sensitive stocks that the *average* investor fears.

Predictable returns

The fact that returns are in fact somewhat predictable modifies the standard portfolio advice in three ways. It introduces horizon effects, it allows market-timing strategies, and it introduces multiple factors via hedging demands (if expected returns vary over time, investors may want to hold assets that protect them against this risk).

Horizon effects

Recall that when stock returns are independent over time (like coin flips), the allocation between stocks and bonds does not depend at all on the investment horizon, since mean returns (reward) and the variance of returns (risk) both increase in proportion to the investment horizon. But if returns are predictable, the mean and variance may no longer scale the same way with horizon. If a high return today implies a high return tomorrow—positive serial correlation—then the variance of returns will increase with horizon faster than does the mean return. In this case, stocks are worse in the long run. If a high return today implies a lower return tomorrow—negative serial correlation or *mean reversion*—then the variance of long-horizon returns is lower than the variance of one-period returns times the horizon. In this case, stocks are more attractive for the long run.[1] For example, if the second coin flip is always the opposite of the first coin flip, then two coin flips are much less risky than they would be if each flip were independent, and a "long-run coin flipper" is more likely to take the bet.

Which case is true? Overall, the evidence suggests that stock prices do tend to come back slowly and partially after a shock, so return variances at horizons of five years and longer are about one-half to two-thirds as large as short-horizon variances suggest. Direct measures of the serial correlation of stock returns, or equivalent direct measures of the mean and variance of long-horizon returns, depend a lot on the period studied and the econometric method. Multivariate methods give somewhat stronger evidence. Intuitively, the price/dividend (p/d) ratio does not explode. Hence, the long-run variance of prices must be the same as the long-run variance of dividends, and this extra piece of information helps to measure the long-run variance of returns. (Cochrane and Sbordone, 1986, and Cochrane, 1994, use this idea. See Campbell, Lo, and MacKinlay, 1997, for a summary of these issues and of the extensive literature.)

How big are the horizon effects? Barberis (1999) calculates optimal portfolios for different horizons when returns are predictable. Figure 3 presents some of his results.

We start with a very simple setup: The investor allocates his portfolio between stocks and bonds and then holds it without rebalancing for the indicated horizon. His objective is to maximize the expected utility of wealth at the indicated horizon. The flat line in figure 3 shows the standard result: If returns are not

predictable, then the allocation to stocks does not depend on horizon.

The top (black) line in figure 3 adds the effects of return predictability on the investment calculation. The optimal allocation to stocks increases sharply with horizon, from about 40 percent for a monthly horizon to 100 percent for a ten-year horizon. To quantify the effects of predictability, Barberis uses a simple model,

$$1) \quad R_{t+1} - R^{TB}_{t+1} = a + bx_t + \varepsilon_{t+1}$$

$$2) \quad x_{t+1} = c + \rho x_t + \delta_{t+1},$$

using the d/p ratio for the forecasting variable x. (Whether or not one includes returns in the right hand side makes little difference.) Barberis estimates significant mean-reversion: In Barberis's regressions, the implied standard deviation of ten-year returns is 23.7 percent, just more than half of the 45.2 percent value implied by the standard deviation of monthly returns. Stocks are indeed safer in the long run, and the greater allocation to stocks shown in figure 3 for a long-run investor reflects this fact.

Uncertainty about predictability

This calculation ignores the fact that we do not know how predictable returns really are. One could address this fact by calculating standard errors for portfolio computations; and such standard errors do indicate substantial uncertainty. However, standard error uncertainty is symmetric—returns might be more predictable than we think or they might be less predictable. This measure of uncertainty would say that we are just as likely to want an even greater long-run stock allocation as we are to shade the advice back to a constant allocation.

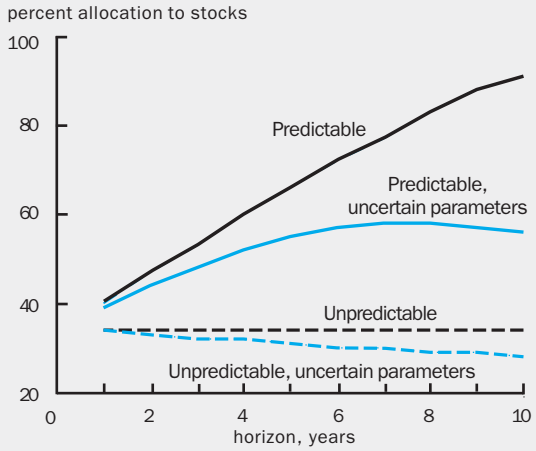
Intuitively, however, uncertainty about predictability should lead us to shade the advice back toward the standard advice. Standard errors do not capture the uncertainties behind this (good) intuition for at least two reasons.

First, the predictability captured in Barberis's regression of returns on dividend/price ratios certainly results to some extent from data-dredging. Thousands of series were examined by many authors, and we have settled on the one or two that seem to predict returns best in sample. The predictability will obviously be worse out of sample, and good portfolio advice should account for this bias. Standard errors take the set of forecasting variables and the functional form as given.

Second, the portfolio calculation assumes that the *investor* knows the return forecasting process perfectly. Standard errors only reflect the fact that *we* do not know the return forecasting process, so we

FIGURE 3

Allocation to stocks, different investment horizons



Notes: The investor maximizes the utility of terminal wealth via a buy-and-hold investment in stocks versus bonds. The investor has constant relative risk aversion utility with a risk aversion coefficient of 10. The top calculation (black) includes predictable returns modeled by a regression on d/p ratios (equation 1). The second calculation (color) includes predictable returns and the effects of parameter uncertainty. The third calculation (black dash) assumes unpredictable returns, and no parameter uncertainty. The bottom calculation (color dash) assumes unpredictable returns, but adds parameter uncertainty. All distributions are conditional on a dividend/price ratio equal to its historical mean.
Source: Barberis (1999).

are unsure about what investors want to do.[2] What we would like to do is to solve a portfolio problem in which investors treat uncertainty about the forecastability of returns as part of the risk that they face, along with the risks represented by the error terms of the statistical model. Kandel and Stambaugh (1996) and Barberis (1999) tackle this important problem.

Figure 3 also gives Barberis's calculations of the effects of parameter uncertainty on the stock/bond allocation problem. (See box 1 for a description of how these calculations are made.) The lowest (dotted) line considers a simple case. The investor knows, correctly, that returns are independent over time (not predictable) but he is not sure about the mean return. Without parameter uncertainty, this situation gives rise to the constant stock allocation—the flat line. Adding parameter uncertainty *lowers* the allocation to stocks for long horizons; it declines from 34 percent to about 28 percent at a ten-year horizon.

The reason is simple. If the investor sees a few good years of returns after making the investment, this raises his estimate of the actual mean return and, thus, his estimate of the returns over the rest of the investment period. Conversely, a few bad years will lower his estimate of the mean return for remaining years. Thus, learning about parameters induces a

BOX 1

How to include model uncertainty in portfolio problems

A statistical model, such as equations 1 and 2, tells us the distribution of future returns once we know the parameters θ, $f(R_{t+1}|\theta, x_1, x_2, ..., x_t)$, where x_t denotes all the data used (returns, d/p, etc.).

We would like to evaluate uncertainty by the distribution of returns conditional only on the history available to make guesses about the future, $f(R_{t+1}|x_1, x_2 \ldots x_t)$. We can use Bayesian analysis to evaluate this concept. If we can summarize the information about parameters given the historical data as $f(\theta|x_1, x_2 \ldots x_t)$, then we can find the distribution we want by

$$f(R_{t+1}|x_1, x_2 \ldots x_t) = \int f(R_{t+1}|\theta) f(\theta|x_1, x_2 \ldots x_t) d\theta.$$

In turn, we can construct $f(\theta|x_1, x_2 \ldots x_t)$, from a prior $f(\theta)$ and the likelihood function $f(x_1, x_2 \ldots x_t|\theta)$ via the standard law for conditional probabilities,

$$f(\theta|x_1, x_2 \ldots x_t) = \frac{f(x_1, x_2 \ldots x_t|\theta) f(\theta)}{f(x_1, x_2 \ldots x_t)}$$

$$f(x_1, x_2 \ldots x_t) = \int f(x_1, x_2 \ldots x_t|\theta) f(\theta) d\theta.$$

Barberis (1999), Kandel and Stambaugh (1996), Brennan, Schwartz, and Lagnado (1997) use these rules to compute $f(R_{t+1}|x_1, x_2 \ldots x_t)$, and solve portfolio problems with this distribution over future returns.

positive correlation between early returns and later returns. Positive correlation makes long-horizon returns more than proportionally risky and reduces the optimal allocation to stocks.

The colored line in figure 3 shows the effects of parameter uncertainty on the investment problem, when we allow return predictability as well. As the figure shows, uncertainty about predictable returns cuts the increase in stock allocation from one to ten years *in half*. In addition to the positive correlation of returns due to learning about their mean mentioned above, uncertainty about the true amount of predictability adds to the risk (including parameter risk) of longer horizon returns.

Market timing

Market-timing strategies are the most obvious implication of return predictability. If there are times when expected returns are high and other times when they are low, you might well want to hold more stocks when expected returns are high, and fewer when expected returns are low. Of course, the crucial question is, how *much* market-timing should you engage in? Several authors have recently addressed this technically challenging question.

Much of the difficulty with return predictability (as with other dynamic portfolio questions) lies in computing the optimal strategy—exactly how should you adjust your portfolio as the return prediction signals change? Gallant, Hansen, and Tauchen (1990) show how to measure the potential benefits of market-timing without actually calculating the market-timing strategy.

The mean–standard deviation tradeoff or *Sharpe ratio*—the slope of the frontier graphed in figure 1—is a convenient summary of any strategy. If the risk-free rate is constant and known, *the square of the maximum unconditional Sharpe ratio is the average of the squared conditional Sharpe ratios*. (The appendix details the calculation.) Since we take an average of *squared* conditional Sharpe ratios, volatility in conditional Sharpe ratios—time-variation in expected returns or return volatility—is good for an investor who cares about the unconditional Sharpe ratio. By moving into stocks in times of high Sharpe ratio and moving out of the market in times of low Sharpe ratio, the investor does better than he would by buying and holding. Furthermore, *the best unconditional Sharpe ratio is directly related to the R^2 in the return-forecasting regression*.

The buy-and-hold Sharpe ratio has been about 0.5 on an annual basis in U.S. data—stocks have earned an average return of about 8 percent over Treasury bills, with a standard deviation of about 16 percent. Table 1 presents a calculation of the increased Sharpe ratio one should be able to achieve by market-timing, based on regressions of returns on dividend/price ratios. (I use the regression estimates from table 1 of "New facts in finance.")

As table 1 indicates, market-timing should be a great benefit. Holding constant the portfolio volatility, market-timing should raise average returns by about two-fifths at an annual horizon and it should almost double average returns at a five-year horizon.

TABLE 1

Maximum unconditional Sharpe ratios

Horizon k (years)	R^2	Annualized Sharpe ratio
Buy & hold		0.50
1	0.17	0.71
2	0.26	0.72
3	0.38	0.78
5	0.59	0.95

Notes: Maximum unconditional Sharpe ratios available from market-timing based on regressions of value-weighted NYSE index returns on the dividend/price ratio. The table reports annualized Sharpe ratios corresponding to each R^2. The formula is $\frac{S^*}{\sqrt{k}} = \sqrt{0.5^2 + \frac{R^2}{k}} / \sqrt{1-R^2}$ and is derived in the appendix.

Optimal market-timing: An Euler equation approach

Brandt (1999) presents a clever way to estimate a market-timing portfolio rule without solving a model. Where standard asset pricing models fix the consumption or wealth process and estimate preference parameters, Brandt fixes the preference parameters (as one does in a portfolio question) and estimates the portfolio decision, that is, he estimates the optimal consumption or wealth process.[3] This calculation is very clever because it does not require one to specify a statistical model for the stock returns (like equations 1–2), and it does not require one to solve the economic model.

Figure 4 presents one of Brandt's results. The figure shows the optimal allocation to stocks as a function of investment horizon and of the dividend/price ratio, which forecasts returns. There is a mild horizon effect, about in line with Barberis's results of figure 3 without parameter uncertainty: Longer term investors hold more stocks. There is also a strong market-timing effect. The fraction of wealth invested in stocks varies by about 200 percentage points for all investors. For example, long-term investors vary from about 75 percent to 225 percent of wealth invested in stocks as the d/p ratio rises from 2.8 percent to 5.5 percent.

Optimal market-timing: A solution

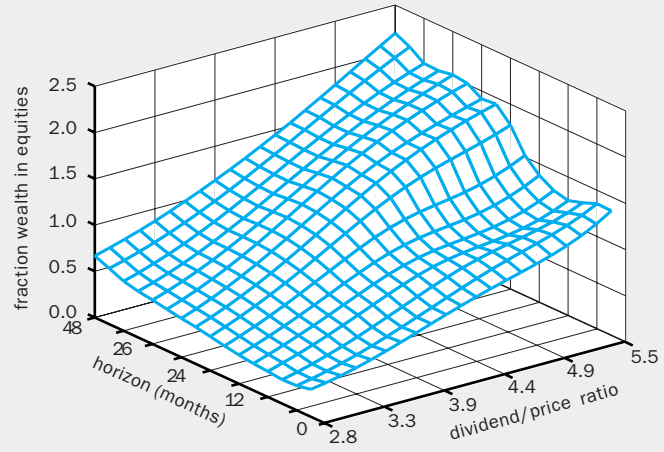
Campbell and Vicera (1999) actually calculate a solution to the optimal market-timing question. They model investors who desire lifetime consumption[4] rather than portfolio returns at a fixed horizon. They model the time-variation in expected stock returns via equation 1 on d/p ratios. Their investors live only off invested wealth and have no labor income or labor income risk. Thus, these investors are poised to take advantage of business cycle related variation in expected returns.

As one might expect, the optimal investment strategy takes strong advantage of market-timing possibilities. Figure 5 reproduces Campbell and Vicera's optimal allocation to stocks as a function of the expected return, forecast from d/p ratios via equation 1. A risk aversion coefficient of 4 implies that investors roughly want to be fully invested in stocks at the average expected excess return of 6 percent, so this is a sensible risk aversion value to consider. Then, as the d/p ratio ranges from minus two to plus one standard deviations from its mean, these investors range from –50 percent in stocks to 220 percent in stocks. This is aggressive market-timing indeed.

Figure 6 presents the calculation in a different way: It gives the optimal allocation to stocks over time, based on dividend/price ratio variation over time. The high d/p ratios of the 1950s suggest a strong stock position, and that strong position profits from the high returns of the late 1950s to early 1960s. The low d/p ratios of the 1960s suggest a much smaller position in stocks, and this smaller position avoids the bad returns of the 1970s. The high d/p ratios of the 1970s suggest strong stock positions again, which benefit from the good return of the 1980s; current

FIGURE 4

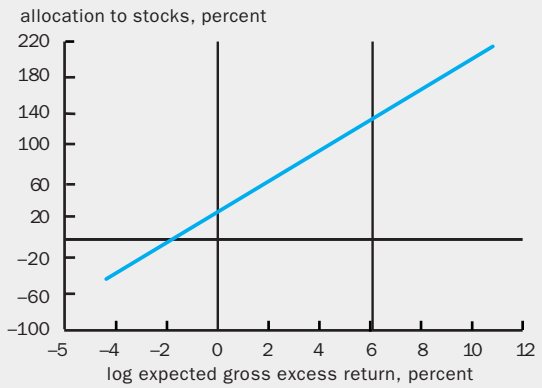
Optimal allocation to stocks



Notes: Optimal allocation to stocks as a function of horizon and dividend yield.
Source: Brandt (1999).

FIGURE 5

Optimal allocation to stocks



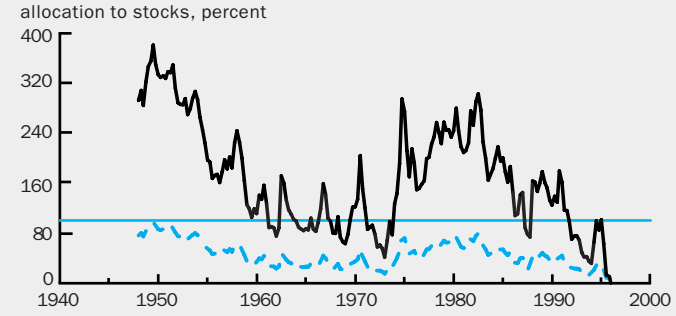
Notes: Optimal allocation to stocks as a function of the expected return implied by a regression that forecasts stock returns from dividend/price ratios. The line extends from a d/p ratio two standard deviations above its mean (low expected returns) to one standard deviation below its mean (high expected returns). Risk aversion is 4.0.
Source: Campbell and Vicera (1999).

unprecedented high prices suggest the lowest stock positions ever. The optimal allocation to stocks again varies wildly, from 0 (now) to over 300 percent.

Campbell and Vicera's calculations are, if anything, conservative compared with others in this literature. Other calculations, using other utility functions, solution techniques, and calibrations of the forecasting process often produce even more aggressive market-timing strategies. For example, Brennan, Schwartz, and Lagnado (1997) make a similar calculation with two additional forecasting variables. They report market-timing strategies that essentially jump back and forth between constraints at 0 percent in stocks and 100 percent in stocks.

FIGURE 6

Optimal allocation to stocks based on dividend/price ratio



Notes: Risk aversion $\gamma = 4.00$ (black line) and $\gamma = 20.00$ (colored dashed line).
Source: Campbell and Vicera (1999).

Campbell and Vicera also present achieved utility calculations that mirror the lesson of table 1: Failing to time the market seems to impose a large cost.

Doubts

One may be understandably reluctant to take on quite such strong market-timing positions as indicated by figures 5 and 6, or to believe table 1 that market timing can nearly double five-year Sharpe ratios. In particular, one might question advice that would have meant missing the dramatic runup in stock values of the late 1990s. Rather than a failure of nerve, perhaps such reluctance reveals that the calculations do not yet include important considerations and, therefore, overstate the desirable amount of market-timing and its benefits.

First, the unconditional Sharpe ratio as reported in table 1 for, say, five-year horizons answers the question, "Over very long periods, if an investor follows the best possible market-timing strategy and evaluates his portfolio based on five-year returns, what Sharpe ratio does he achieve?" It does *not* answer the question, "Given today's d/p, what is the best Sharpe ratio you can achieve for the next five years by following market-timing signals?" The latter question characterizes the return distribution conditional on today's d/p. It is harder to evaluate; it depends on the initial d/p level, and it is lower, especially for a slow-moving signal such as d/p.

To see the point, suppose that the d/p ratio is determined on day one, is constant thereafter, and indicates high or low returns in perpetuity. *Conditional* on the d/p ratio, one cannot time the market at all. But since the investor will invest less in stocks in the low-return state and more in the high-return state, he will *unconditionally* time the market (that is, adjust his portfolio based on day one information) and this gives him a better date-zero (unconditional) Sharpe ratio than he would obtain by fixing his allocation at date zero. This fact captures the intuition that there is a lot more money to be made from a 50 percent R^2 at a daily horizon than at a five-year horizon, where the calculations in table 1 are not affected by the persistence of the market-timing signal. Campbell and Vicera's (1999) utility calculations are also based on the unconditional distribution, so the optimal degree and benefit of market-timing might be less, conditional on the observed d/p ratio at the first date.

Second, there are good statistical reasons to think that the regressions overstate the predictability of returns.

1) Figure 6 emphasizes one reason: The d/p ratio signal has only crossed its mean four times in the 50 years of postwar history. You have to be very patient to profit from this trading rule. Also, we really have only four postwar data points on the phenomenon. 2) The dividend/price ratio was selected, in sample, among hundreds of potential forecasting variables. It has not worked well out of sample—the last two years of high market returns with low d/p ratios have cut the estimated predictability in half! 3) The model imposes a linear specification, where the actual predictability is undoubtedly better modeled by some unknown nonlinear function. In particular, the linear specification implies negative expected stock returns at many points in the sample, and one might not want to take this specification seriously for portfolio construction. 4) The d/p ratio is strongly autocorrelated, and estimates of this autocorrelation are subject to econometric problems. For this reason, long-horizon return properties inferred from a regression such as equations 1 and 2 are often more dramatic and apparently more precisely measured than direct long-horizon estimates.

The natural next step is to include this parameter uncertainty in the portfolio problem, as I did above for the case of independent returns. While this has not been done yet in a model with Campbell and Vicera's (1999) level of realism (and for good reasons—Campbell and Vicera's non-Bayesian solution is already a technical tour de force), Barberis (1999) makes such calculations in his simpler formulation. He uses a utility of terminal wealth and no intermediate trading, and he forces the allocation to stocks to be less than 100 percent.

Figure 7 presents Barberis's (1999) results.[5] As the figure shows, uncertainty about the parameters of the regression of returns on d/p almost eliminates the usefulness of market-timing.

Third, it is uncomfortable to note that fund returns still cluster around the (buy-and-hold) market Sharpe ratio (see figure 7 of "New facts in finance"). Here is a mechanical strategy that supposedly earns average returns twice those of the market with no increase in risk. If the strategy is real and implementable, one must argue that funds simply failed to follow it.

Market-timing, like value, does requires patience and the willingness to stick with a portfolio that departs from the indexing crowd. For example, a market-timer following Campbell and Vicera's rules in figures 5 and 6 would have missed most of the great runup in stocks of the last few years. Fund managers who did that are now unemployed. On the other hand, if an eventual crash comes, the market timer will look wise.

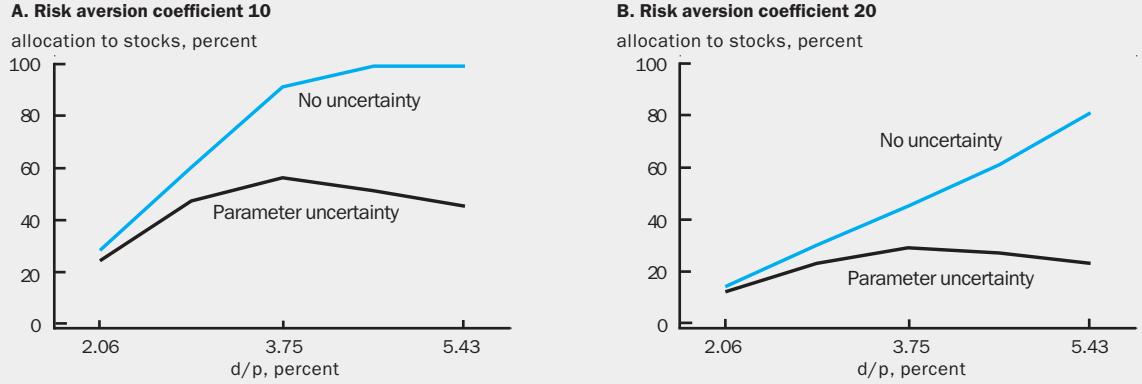
Finally, one's reluctance to take such strong market-timing advice reflects the inescapable fact that getting more return requires taking on more, or different, kinds of risk. A market-timer must buy at the bottom, when everyone else is in a panic; he must sell at the top (now) when everyone else is feeling flush. His portfolio will have a greater mean for a given level of variance over very long horizons, but it will do well and badly at very different times from everyone else's portfolio. He will often underperform a benchmark.

Hedging demands

Market-timing addresses whether you should *change* your allocation to stocks over time as a return signal rises or falls. Hedging demands address whether your *overall* allocation to stocks, or to specific portfolios, should be higher or lower as a result of

FIGURE 7

Allocation to stocks as a function of dividend/price ratio, with parameter uncertainty



Notes: The colored line ignores parameter uncertainty, as in Campbell and Vicera (1999).
The black line includes parameter uncertainty, as in Barberis (1999). Data sample is in months (523).

return predictability, in order to protect you against reinvestment risk.

A long-term bond is the simplest example. Suppose you want to minimize the risk of your portfolio ten years out. If you invest in apparently safe short-term risk-free assets like Treasury bills or a money-market fund, your ten-year return is in fact quite risky, since interest rates can fluctuate. You should hold a ten-year (real, discount) bond. Its price will fluctuate a lot as interest rates go up and down, but its value in ten years never changes.

Another way of looking at this situation is that, if interest rates decline, the price of the ten-year bond will skyrocket; it will skyrocket just enough so that, reinvested at the new lower rates, it provides the same ten-year return as it would have if interest rates had not changed. Changes in the ten-year bond value *hedge* the reinvestment risk of short-term bonds. If lots of investors want to secure the ten-year value of their portfolios, this will raise demand for ten-year bonds and lower their prices.

In general, the size and sign of a hedging demand depend on risk aversion and horizon and, thus, will be different for different investors. If the investor is quite risk averse—infinitely so in my bond example—he wants to buy assets whose prices go up when expected returns decline. But an investor who is not so risk averse might want to buy assets whose prices go up when expected returns *rise*. If the investor is sitting around waiting for a good time to invest, and is willing to pounce on good (high expected return) investments, he would prefer to have a lot of money to invest when the good opportunity comes around. It turns out that the dividing line in the standard (CRRA) model is logarithmic utility or a risk aversion coefficient of 1—investors more risk averse than this want assets whose prices go up when expected returns decline, and vice versa. Most investors are undoubtedly more risk averse than this, but not necessarily all investors. Horizon matters as well. A short horizon investor cares nothing about reinvestment risk and, therefore, has zero hedging demand.

In addition, the relationship between price and expected returns is not so simple for stocks as for bonds and must be estimated statistically. The predictability evidence reviewed above suggests that high stock returns presage lower subsequent returns. High returns drive up price/dividend, price/earnings, and market/book ratios, all of which have been strong signals of lower subsequent returns. Therefore, stocks are a good hedge against their own reinvestment risk—they act like the long-term assets that they are. This consideration raises the attractiveness of stocks for typical (risk aversion greater than 1) investors. Precisely, if the two-fund analysis of figure 1 suggests a certain split between stocks and short-term bonds for a given level of risk aversion and investment horizon, then return predictability, a long horizon, and typical risk aversion greater than 1 will result in a higher fraction devoted to stocks. Again, exactly how *much* more one should put into stocks in view of this consideration is a tough question.

(In this case, the hedging demand reduces to much the same logic as the horizon effects described above. The market portfolio is a good hedge against its own reinvestment risk, and so its long horizon variance is less than its short horizon variance would suggest. More generally, hedging demands can tilt a portfolio toward stocks whose returns better predict and, hence, better hedge the expected return on the market index, but this long-studied possibility from Merton [1971a, 1971b] has not yet been implemented in practice.)

Campbell and Vicera's (1999) calculations address this hedging demand as well as market-timing demand, and figure 5 also illustrates the strength of the hedging demand for stocks. Campbell and Vicera's investors want to hold almost 30 percent of their wealth in stocks even if the expected return of stocks is no greater than that of bonds. Absent the hedging motive, of course, the optimal allocation to stocks would be zero with no expected return premium. Almost a 2 percent *negative* stock return premium is necessary to dissuade Campbell and Vicera's investors from holding stocks. At the average (roughly 6 percent) expected return, of the roughly 130 percent of wealth that the risk aversion 4 investors want to allocate to stocks, nearly half is due to hedging demand. Thus, hedging demands can importantly change the allocation to stocks.

However, hedging demand works in opposition to the usual effects of risk aversion. Usually, less risk averse people want to hold more stocks. However, less risk averse people have lower or even negative hedging demands, as explained above. It is possible that hedging demand exactly offsets risk aversion; everybody holds the same mean allocation to stocks. This turns out not to be the case for Campbell and Vicera's numerical calibration; less risk averse people still allocate more to stocks on average, but the effect depends on the precise specification.

Choosing a risk-free rate

Figure 1 describes a portfolio composed of the market portfolio and the risk-free rate. But the risk-free rate is not as simple as it once was either. For a consumer or an institution[6] with a one-year horizon,

one-year bonds are risk-free, while for one with a ten-year horizon, a ten year zero-coupon bond is risk-free. For a typical consumer, whose objective is lifetime consumption, an interest-only strip (or real level annuity) is in fact the risk-free rate, since it provides a riskless coupon that can be consumed at each date. Campbell and Vicera (1998) emphasize this point. Thus, the appropriate bond portfolio to mix with risky stocks in the logic of figure 1 is no longer so simple as a short-term money market fund.

Of course, these comments refer to *real* or indexed bonds, which are only starting to become easily available. When only nominal bonds are available, the closest approximation to a risk-free investment depends additionally on how much interest rate variability is due to real rates versus nominal rates. In the extreme case, if real interest rates are constant and nominal interest rates vary with inflation, then rolling over short-term nominal bonds carries less long-term real risk than holding long-term nominal bonds. In the past, inflation was much more variable than real interest rates in the U.S., so the fact that portfolio advice paid little attention to the appropriate risk-free rate may have made sense. We seem to be entering a period in which inflation is quite stable, so *real* interest rate fluctuations may dominate interest rate movements. In this case, longer term nominal bonds become more risk-free for long-term investors, and inflation-indexed bonds open up the issue in any case. Once again, new facts are opening up new challenges and opportunities for portfolio formation.

Notes of caution

The new portfolio theory can justify all sorts of interesting new investment approaches. However, there are several important qualifications that should temper one's enthusiasm and that shade portfolio advice back to the traditional view captured in figure 1.

The average investor holds the market

The portfolio theory that I have surveyed so far asks, given multiple factors or time-varying investment opportunities, How should an investor *who does not care* about these extra risks profit from them? This may result from intellectual habit, as the past great successes of portfolio theory addressed such investors, or it may come from experience in the money management industry, where distressingly few investors ask about additional sources of risk that multifactor models and predictable returns suggest should be a major concern.

Bear in mind, however, that *the average investor must hold the market portfolio.* Thus, *multiple factors and return predictability cannot have any portfolio implications for the average investor*. In addition, for every investor who should follow a value strategy or time the market for the extra returns offered by those extra risks, *there must be an investor who should follow the exact opposite advice*. He should follow a growth strategy or sell stocks at the bottom and buy at the top, because he is unusually exposed to or averse to the risks of the value or market-timing strategies in his business or job. He knows that he pays a premium for not holding those risks, but he rationally chooses this course just as we all choose to pay a premium for home insurance.

Again, dividend/price, price/earnings, and book/market ratios forecast returns, if they do, *because* the average investor is unwilling to follow the value and market-timing strategies. If everyone tries to time the market or buy more value stocks, the premiums from these strategies will disappear and the CAPM, random walk view of the market will reemerge. Market-timing can only work if it involves buying stocks when nobody else wants them and selling them when everybody else wants them. Value and small-cap anomalies can only work if the average investor is leery about buying financially distressed and illiquid stocks. Portfolio advice to follow these strategies *must* fall on deaf ears for the average investor, and a large class of investors must want to head in exactly the other direction. If not, the premiums from these strategies will not persist.

One can see a social function in all this: *The stock market acts as a big insurance market*. By changing weights in, say, recession-sensitive stocks, people whose incomes are particularly hurt by recessions can purchase insurance against that loss from people whose incomes are not hurt by recessions. They pay a premium to do so, which is why investors are willing to take on the recession-related risk.

The quantitative portfolio advice is all aimed at the providers of insurance, which may make sense if the providers are large wealthy investors or institutions. But for each provider of insurance, there must be a purchaser, and his portfolio must take on the opposite characteristics.

Are the effects real or behavioral, and will they last?

So far, I have emphasized the view that the average returns from multifactor or market-timing strategies are earned as compensation for holding real, aggregate risks that the average investor is anxious not to hold. This view is still debated. Roughly half of the academic studies that document such strategies interpret them this way, while the other half interpret them as evidence that investors are systematically irrational. This half argues that a new "behavioral finance"

should eliminate the assumption of rational consumers and investors that has been at the core of all economics since Adam Smith, in order to explain these asset pricing anomalies.

For example, I have followed Fama and French's (1993, 1996) interpretation that the *value effect* exposes the investor to systematic risks associated with economywide financial distress. However, Lakonishok, Shleifer, and Vishny (1994) interpret the same facts as evidence for irrationality: Investors flock to popular stocks and away from unpopular stocks. The prices of the unpopular stocks are depressed, and their average returns are higher as the fad slowly fades. Fama and French point out that the behavioral view cannot easily account for the comovement of value stocks; the behavioral camp points out that the fundamental risk factor is still not determined.

Similarly, the predictability of stock returns over time is interpreted as waves of irrational exuberance and pessimism as often as it is interpreted as time-varying, business cycle related risk or risk aversion. Those who advocate an economic interpretation point to the association with business cycles (Fama and French, 1989) and to some success for explicit models of this association (Campbell and Cochrane, 1999); those who favor the irrational investors view point out that the rational models are as yet imperfect.

While this academic debate is entertaining, how does it affect a practical investor who is making a portfolio decision? At a basic level, it does not. If *you* are not exposed to the risk a certain investment represents, it does not matter why other investors shy away from holding it.

Analogously, to decide what to buy at the grocery store, you only have to know how you feel about various foods and what their prices are. You do not have to understand the economic determinants of food prices: You do not need to know whether a sale on tomatoes represents a "real" factor like good weather in tomato growing areas, or whether it represents an "irrational" fear or fad.

Will they last?

Investments do not come with average returns as clearly marked as grocery prices, however. Investors have to figure out whether an investment opportunity that did well in the past will continue to do well. This is one reason that it is important to understand whether average returns come from real or irrational aversion to risk.

If it is *real*, it is most likely to persist. If a high average return comes from exposure to risk, well understood and widely shared, that means all investors understand the opportunity but shrink from it. Even if the opportunity is widely publicized, investors will not change their portfolio decisions, and the relatively high average return will remain.

On the other hand, if it is truly *irrational*, or a market inefficiency, it is least likely to persist. If a high average return strategy involves no extra exposure to real risks and is easy to implement (it does not incur large transaction costs), that means that the average investor will immediately want to invest when he hears of the opportunity. News travels quickly, investors react quickly, and such opportunities vanish quickly.

Recent work in behavioral finance tries to document a way that irrational phenomena can persist in the face of the above logic. If an asset-pricing anomaly corresponds to a fundamental, documented, deeply formed aspect of human psychology, then the average investor may *not* pounce on the strategy the minute he hears of it, and the phenomenon may last (DeBondt and Thaler, 1985, and Daniel, Hirshleifer, and Subrahmanyam, 1998). For example, many people systematically overestimate the probability that airplanes crash, and make wrong decisions resulting from this belief, such as choosing to drive instead. No amount of statistics changes this view. Most such people readily admit that a fear of flying is "irrational" but persist in it anyway. If an asset-pricing anomaly results from such a deep-seated misperception of risk, then it could in fact persist.

A final possibility is that the average return premiums are the result of *narrowly held risks.* This view is (so far) the least stressed in academic analysis. In my opinion, it may end up being the most important. It leads to a view that the premiums will be moderately persistent. Catastrophe-insurance enhanced bonds provide a good example of this effect. These bonds pay well in normal times, but either part of the principal or interest is pledged against a tranche of a property reinsurance contract. Thus, the bonds promise an average return of 10 percent to 20 percent (depending on one's view of the chance of hurricanes). However, the risk of hurricane damage is uncorrelated with anything else, and hence it is perfectly diversifiable. Therefore, catastrophe bonds are an attractive opportunity. Before the introduction of catastrophe bonds, there was no easy way for the average investor or fund to participate in property reinsurance. As more and more investors and funds hold these securities, the prices will rise and average returns will fall. Once the risks are widely shared, every investor (at least those not located in hurricane-prone areas) will hold a little bit of the risk and the high average returns will have vanished.

The essential ingredients for this story are that the risk is narrowly shared; the high average returns only disappear when the risk is widely shared (it cannot be arbitraged away by a few savvy investors); and an institutional change (the introduction, packaging, and marketing of catastrophe-linked bonds) is required before it all can happen.

This story gives a plausible interpretation of many of the anomalies I document above. Small-cap stocks were found in about 1979 to provide higher returns than was justified by their market (β) risk. Yet at that time, most funds did not invest in such stocks, and individual investors would have had a hard time forming a portfolio of small-cap stocks without losing all the benefits in the very illiquid markets for these stocks. The risks were narrowly held. After the popularization of the small-cap effect, many small-cap funds were started, and it is now easy for investors to hold such stocks. As the risk has been more widely shared, the average returns seem to have fallen.

The value effect may be amenable to a similar interpretation. Before about 1990, as I noted earlier, few funds actually followed the high-return strategy of buying really distressed stocks or shorting the popular growth stocks. It would be a difficult strategy for an individual investor to follow, requiring courage and frequent trading of small illiquid stocks. Now that the effect is clear, value funds have emerged that really do follow the strategy, and the average investor can easily include such an offering in his portfolio. The risk is becoming widely shared, and its average return seems to be falling.

Even average returns on the stock market as a whole (the equity premium) may follow the same story, since participation has increased a great deal through the invention of index funds, low-commission brokerages, and tax-sheltered retirement plans.

This story does not mean that the average returns corresponding to such risks will vanish. They will decline, however, until the markets have established an equilibrium, in which every investor has bought as much of the risk as he likes. In this story, one would expect a large return as investors discover each strategy and bid prices up to their equilibrium levels. This may account for some of the success of small and value stocks observed in the literature, as well as some of the stunning success of the overall market in recent years.

Inconsistent advice

Unfortunately, *the arguments that a factor will persist are all inconsistent with aggressive portfolio advice*. If the premium is *real*, an equilibrium reward for holding risk, then the average investor knows about it but does not invest because the extra risk exactly counteracts the extra average return. If more than a minuscule fraction of investors are not already at their best allocations, then the market has not reached equilibrium and the premiums will change.

If the risk is *irrational*, then by the time you and I know about it, it's gone. An expected return corresponding to an irrational risk premium has the strongest portfolio implications—everyone should do it—but the shortest lifetime. Thus, this view is also inconsistent with the widespread usefulness of portfolio advice.

If the average return comes from a *behavioral* aversion to risk, it is just as inconsistent with widespread portfolio advice as if were real. We can not all be less behavioral than average, just as we can not all be less exposed to a risk than average. The whole argument for behavioral persistence is that the average investor would not change his portfolio, just as the average traveler would not quickly adjust his traveling behavior to fear the cab ride out to the airport more than the flight. Thus, the advice *must* be useless to the vast majority of investors. If most people, on seeing the strategy, can be persuaded to act differently and buy, then it is an irrational risk and will disappear. If it is real or behavioral and will persist, then this *necessarily* means that very few people will follow the portfolio advice.

If the average return comes from a *narrowly held* risk, one has to ask what institutional barriers keep investors from sharing this risk more widely. Simple portfolio advice may help a bit—most investors still do not appreciate the risk/return advantages of stocks overall, small-caps, value stocks, market-timing, and aggressive liquidity trades. But by and large, a risk like this needs packaging, securitizing, and marketing more than advice. Then there will be a period of high average returns to the early investors, followed by lower returns, but still commoditization of the product with fees for the intermediaries.

Economic logic

The issue of why the risk gives an average return premium is also important to decide whether the opportunity is really there. It is not that easy to establish the average returns of stocks and dynamic portfolio strategies. There are many statistical anomalies that vanish quickly out of sample. Figuring out *why* a strategy carries a high average return is one of the best ways to ensure that the high average return is really there in the first place. Anything that is going to work has a real economic function. A story such

as "I don't care much about recessions; the average investor does; hence it makes good sense for me to buy extra amounts of recession sensitive stocks since I am selling insurance to the others at a premium" makes a strategy much more plausible than the output of some statistical black box.

Conclusion

Practical application of portfolio theory

How does an investor who is trying patiently to sort through the bewildering variety of investment opportunities use all the new portfolio theory? It's best to follow a step by step procedure, starting with a little introspection.

1. *What is your overall risk tolerance?* As before, you must first figure out to what extent you are willing to trade off volatility for extra average returns, to determine an appropriate overall allocation to risky versus risk-free assets. While this question is hard to answer in the abstract, you only need to know whether you are more or less risk tolerant than the average investor. (Honestly, now—everyone wants to say they are a risk taker.) The overall market is about 60 percent stocks and 40 percent bonds, so average levels of risk aversion, whatever they are, wind up at this value.

2. *What is your horizon?* This question is first of all important for figuring out what is the relevant risk-free asset. Longer term investors can hold longer term bonds despite their poor one-year performance, especially in a low-inflation environment. Second, we have seen that stocks are somewhat safer for "long-run" investors.

3. *What are your risks?* Would you be willing to trade some average return in order to make sure that your portfolio does well in particular circumstances? For example, an investor who owns a small company would not want his investment portfolio to do poorly at the same time that his industry suffers a downturn, that there is a recession, or a credit crunch, or that the industries he sells to suffer a downturn. Thus, it makes good sense for him to avoid stocks in the same industry or downstream industries, or stocks that are particularly sensitive to recessions or credit crunches, or even to short them if possible. This strategy would make sense even if these stocks give high average returns, like the value portfolios. Similarly, he should avoid high yield bonds that will all do badly in a credit crunch. If the company will do poorly in response to increases in interest rates, oil prices or similar events, and if the company does not hedge these risks, then the investor should take positions in interest rate sensitive or oil-price sensitive securities to offset those risks as well. We're just extending the principles behind fire and casualty insurance to investment portfolios.

This logic extends beyond the kind of factors (size, book to market, and so on) that have attracted academic attention. It applies to any identifiable movement in asset portfolios. For example, industry portfolios are not badly explained by the CAPM, as they all seem to have about the same average return. Therefore, they do not show up in multifactor models. However, shorting your industry portfolio protects you against the risks of your occupation. In fact, factors that do *not* carry unusual risk premiums are even better opportunities than the priced factors that attract attention, since you buy insurance at zero premium. This was always true, even in the CAPM, unpredictable return view. I think that the experience with multifactor models just increases our awareness of how important this issue is.

4. *What are **not** your risks?* Next, figure out what risks you do not face, but that give rise to an average return premium in the market because most other investors do face these risks. For example, an investor who has no other source of income beyond his investment portfolio does not particularly care about recessions. Therefore, he should buy extra amounts of recession-sensitive stocks, value stocks, high yield bonds, etc., if these strategies carry a credible high average return. This action works just like selling insurance, in return for a premium. This is the type of investor for whom all the portfolio advice is well worked out.

In my opinion, too many investors think they are in this class. The extra factors and time-varying returns would not be there (and will quickly disappear in the future) if lots of people were willing and able to take them. The presence of multiple factors wakes us up to the possibility that we, like the average investor, may be exposed to extra risks, possibly without realizing it.

5. *Apply the logic of the multifactor-efficient frontier.* Figure 2 now summarizes the basic advice. After thinking through which risk factors are good to hold, and which ones you are already too exposed to; after thinking through what extra premiums you are likely to get for taking on extra risks, you can come to a sensible decision about which risks to take and which to hedge.

6. *Do not forget, the average investor holds the market*. If you're pretty much average, all this thought will lead you right back to holding the market index. To rationalize anything but the market portfolio, you have to be different from the average investor in some identifiable way. The average investor sees some risk in value stocks that counteracts their attractive average returns. Maybe you should too! Right now the average

investor is feeling very wealthy and risk-tolerant, therefore stock prices have risen to unprecedented levels and expected stock returns look very low. It's tempting to sell, but perhaps you're feeling pretty wealthy and risk-tolerant as well.

7. *Of course, avoid taxes and snake oil*. The marketing of many securities and funds is not particularly clear on the nature of the risks. There is no reliable extra return without risk. The economic reasoning in this article should be useful to figure out exactly what type of risk a specific fund or strategy is exposed to, and then whether it is appropriate for you. The average actively managed fund still underperforms its style benchmark, and past performance has almost no information about future performance.

The most important piece in traditional portfolio advice applies as much as ever: A*void taxes and transaction costs*. The losses from churning a portfolio and paying needless short-term capital gain, inheritance, and other taxes are larger than any of the multifactor and predictability effects I have reviewed. Tax issues are much less fun but more important to the bottom line.

A big insurance market

It is tempting to think of asset markets like a racetrack, but they are in reality a big insurance market. Value funds seem to provide extra returns to their investors by buying distressed stocks on the edge of bankruptcy. Long-Term Capital Management was, it seems, providing catastrophe insurance by intermediating liquid assets that investors like into illiquid assets that were vulnerable to a liquidity crunch. Who better to provide catastrophe insurance than rich investors with no other labor income or other risk exposure? Once again, we are reminded that Adam Smith's invisible hand guides self-interested decisions to socially useful ends, often in mysterious ways.

However, asset markets could be better insurance markets. Both new and old portfolio advice implies that the typical investor should hold a stock position that is *short* his company, industry, or other easily hedgeable kinds of risk. Many managers and some senior employees must hold long positions in their own companies, for obvious incentive reasons. But there is no reason that this applies to union pension funds, for example. A little marketing and help from policy should make funds that hedge industry-specific risks to labor income much more attractive vehicles.

APPENDIX

Multifactor portfolio mathematics

This section summarizes algebra in Fama (1996). The big picture is that we still get a hyperbolic region since betas are linear functions of portfolio weights just like means.

The problem is, minimize the variance of a portfolio given a value for the portfolio mean and its beta on some factor. Let

$$w = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_N \end{bmatrix}; R = \begin{bmatrix} R^1 \\ R^2 \\ \vdots \\ R^N \end{bmatrix}; 1 = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}; \beta = \begin{bmatrix} \beta_{1,F} \\ \beta_{2,F} \\ \vdots \\ \beta_{N,F} \end{bmatrix}.$$

Then the portfolio return is

$$R^p = w'R;$$

the condition that the weights add up to 1 is

$$1 = 1'w.$$

The mean of the portfolio return is

$$E(R^p) = E(w'R) = w'E(R) = w'E.$$

The last equality just simplifies notation. The beta of the portfolio on the extra factor is

$$\beta^p = w'\beta.$$

The variance of the portfolio return is

$$\text{var}(R^p) = w'Vw,$$

where V is the variance-covariance matrix of returns. The problem is then

$$\min_w \frac{1}{2} w'Vw \text{ s.t. } w'E = \mu;\ w'1 = 1;\ w'\beta = \beta^p.$$

The Lagrangian is

$$\frac{1}{2} w'Vw - \lambda_0(w'E - \mu) - \lambda_1(w'1 - 1) - \lambda_2\left(w'\beta - \beta^p\right).$$

The first order conditions with respect to w give

$$w = V^{-1}\left(E\lambda_0 + 1\lambda_1 + \beta\lambda_2\right) = V^{-1}A\lambda$$

where

$$A = \left[E \;\; 1 \;\; \beta \right]$$
$$\lambda = \left[\lambda_0 \;\; \lambda_1 \;\; \lambda_2 \right]'$$
$$\delta = \left[\mu \;\; 1 \;\; \beta^p \right]'.$$

Plugging this value of w into the constraint equations

$$A'w = \delta,$$

we get

$$AV^{-1}A\lambda = \delta$$
$$\lambda = (A'V^{-1}A)^{-1}\delta$$
$$w = V^{-1}A(A'V^{-1}A)^{-1}\delta.$$

The portfolio variance is then

$$\text{var}(R^p) = w'Vw = \delta'(A'V^{-1}A)^{-1}\delta.$$

Or, writing out the sum of the matrix notation,

$$\text{Var}(R^p) = \left[\mu \;\; 1 \;\; \beta^p \right] (A'V^{-1}A)^{-1} \begin{bmatrix} \mu \\ 1 \\ \beta^p \end{bmatrix}.$$

The variance is a quadratic function of the mean return and of the desired beta on additional factors. That's why we draw cup-shaped frontiers. As with the mean-variance case, the multifactor efficient frontier is a revolution of a hyperbola. If V is a second moment matrix, to handle a risk-free rate,

$$\text{var}(R^p) = \delta'(A'V^{-1}A)^{-1}A'V^{-1}\Sigma V^{-1}A(A'V^{-1}A)^{-1}\delta,$$

where Σ now denotes the return variance-covariance matrix.

Finding the benefits of a market timing strategy without computing the strategy

I show first that the squared maximum unconditional Sharpe ratio is the average of the squared conditional Sharpe ratios when the riskfree rate is constant,

$$s^{*2} = E\left[s_t^2 \right],$$

where

$$s^* = \max E(R - R^f) / \sigma(R - R^f)$$

denotes the unconditional Sharpe ratio, and

$$s_t = \max E_t(R - R^f) / \sigma_t(R - R^f)$$

denotes the conditional Sharpe ratio.

The technique exploits ideas from Gallant, Hansen, and Tauchen (1990). I exploit Hansen and Jagannathan's (1991) theorem that for any excess return Z and discount factor m such that $0 = E(mZ)$, we have

$$\frac{E(Z)}{\sigma(Z)} \le \frac{\sigma(m)}{E(m)},$$

and equality is attained for some choice of m. Thus, the maximal unconditional Sharpe ratio is

$$\max\left(\frac{E}{\sigma}\right) = \frac{\sigma(m^*)}{E(m^*)},$$

where m^* solves

$$m^* = \arg\min_{\{m\}} \sigma(m) \text{ s.t.}$$
$$E_t(m_{t+1}Z_{t+1}) = 0;\; E_t(m_{t+1}) = 1/R_t^f.$$

Gallant, Hansen, and Tauchen show how to solve this problem in quite general situations. They phrase their result as a "lower bound on discount factor volatility" but given $E(Z)/\sigma(Z) \le \sigma(m)/E(m)$, one can read the maximum slope of the unconditional mean-variance frontier (Sharpe ratio) available from market-timing portfolios. To keep the calculation transparent and simple, I specialize to the case of a constant and observed real risk-free rate $R^f = 1/E_t(m)$. Then, *the unconditional squared Sharpe ratio is the average of the conditional squared Sharpe ratios*,

$$\frac{\sigma^2(m)}{E(m)^2} = \frac{\sigma[E_t(m)^2] + E[\sigma_t^2(m)]}{E(m)^2} = E\left(\frac{\sigma_t^2(m)}{E_t(m)^2}\right).$$

Next, I show that when we forecast stock returns with a regression such as equation 1, and interest rates and the conditional variance of the error term are constant, then the best unconditional Sharpe ratio is related to the regression R^2 by

$$s^* = \frac{\sqrt{s_0^2 + R^2}}{\sqrt{1 - R^2}},$$

where $s_0 = E(R - R^f) / \sigma(R - R^f)$ denotes the unconditional buy-and-hold Sharpe ratio.

If the conditional Sharpe ratio is generated by a single asset (the market), and a linear model with constant error variance,

$$Z_{t+1} = EZ + b(x_t - Ex) + \varepsilon_{t+1},$$

then,

$$\left(\frac{\sigma_t(m)}{E_t(m)}\right)^2 = \left(\frac{E_t(Z)}{\sigma_t(Z)}\right)^2 = \left(\frac{EZ + b(x_t - Ex)}{\sigma_\varepsilon}\right)^2,$$

and

$$E\left(\frac{\sigma_t^2(m)}{E_t(m)^2}\right) = E\left[\left(\frac{EZ + b(x_t - Ex)}{\sigma_\varepsilon}\right)^2\right]$$

$$= \frac{(E(Z))^2 + b^2\sigma^2(x)}{\sigma_\varepsilon^2}$$

$$= \frac{(E(Z))^2}{(1-R^2)\sigma^2(Z)} + \frac{b^2\sigma^2(x)}{(1-R^2)\sigma^2(Z)}$$

$$= \frac{1}{(1-R^2)}\frac{(EZ)^2}{\sigma^2(Z)} + \frac{R^2}{(1-R^2)}$$

$$= \frac{1}{(1-R^2)}\left[\left(\frac{E(Z)}{\sigma(Z)}\right)^2 + R^2\right].$$

The last line demonstrates $s^* = \sqrt{s_0^2 + R^2}/\sqrt{1-R^2}$. To obtain the annualized Sharpe ratios reported in table 1, I divide by the square root of horizon, since mean returns roughly scale with horizon and standard deviations roughly scale with the square root of horizon.

NOTES

[1]To be precise, these statements refer to the conditional serial correlation of returns. It is possible for the conditional serial correlations to be non-zero, resulting in conditional variances that increase with horizon faster or slower than linearly, while the unconditional serial correlation of returns is zero. Conditional distributions drive portfolio decisions.

[2]This effort falls in a broader inquiry in economics. Once we recognize that people are unlikely to have much more data and experience than economists, we have to think about economic models in which people *learn* about the world they live in through time, rather than models in which people have so much history that they have learned all there is to know about the world. See Sargent (1993) for a review of learning in macroeconomics.

[3]The standard first-order condition for optimal consumption and portfolio choice is

$$3) \quad E[(c_{t+1})^{-\gamma} Z_{t+1}] = 0,$$

where c denotes consumption, Z denotes an excess return, and γ is a preference parameter. We usually take data on c and Z, estimate γ, and then test whether the condition actually does hold across assets. In a portfolio problem, however, we *know* the preference parameter γ, but we want to estimate the portfolio. For example, in the simplest case of a one-period investment problem, consumption equals terminal wealth. Equation 3 then becomes

$$E[(\alpha R^f + (1-\alpha) R^m_{t+1})^{-\gamma} Z_{t+1}] = 0.$$

Brandt uses this condition to estimate the portfolio allocation α. He extends the technique to multiperiod problems and problems in which the allocation decision depends on a forecasting variable, that is, market-timing problems.

[4]That is, Campbell and Vicera model investors' objectives by a utility function, max $E\sum_t \beta^t u(c_t)$ rather than a desire for wealth at some particular date, max $Eu(W_T)$.

[5]I thank Nick Barberis for providing this figure. While it is not in Barberis (1999), it can be constructed from results given in that paper.

[6]Of course, in theory, institutions, as such, should not have preferences, as their stockholders or residual claimants can unwind any portfolio decisions they make. This is the famous Modigliani-Miller theorem. In practice, institutions often make portfolio decisions as if they were individuals, and people purveying portfolio advice will run into many such institutions.

REFERENCES

Barberis, Nicholas, 1999, "Investing for the long run when returns are predictable," *Journal of Finance*, forthcoming.

Black, Fischer, and Robert Litterman, 1991, "Global asset allocation with equities, bonds, and currencies," *Goldman Sachs Fixed Income Research.*

Brandt, Michael W., 1999, "Estimating portfolio and consumption choice: A conditional Euler equations approach," *Journal of Finance*, October, forthcoming.

Brennan, Michael J., Eduardo S. Schwartz, and Roland Lagnado, 1997, "Strategic asset allocation," *Journal of Economic Dynamics and Control*, Vol. 21, No. 7, pp. 1377–1403.

Campbell, John Y., and John H. Cochrane, 1999, "By force of habit: A consumption-based explanation of aggregate stock market behavior," *Journal of Political Economy*, Vol. 107, No. 2, April, pp. 205–251.

Campbell, John Y., Andrew W. Lo, and A. Craig MacKinlay, 1996, *The Econometrics of Financial Markets* Princeton, NJ: Princeton University Press.

Campbell, John Y. and Luis M. Vicera, 1999, "Consumption and portfolio decisions when expected returns are time varying," *Quarterly Journal of Economics*, forthcoming.

_____, 1998, "Who should buy long term bonds?," Harvard University, manuscript.

Cochrane, John H., and Argia M. Sbordone, 1988, "Multivariate estimates of the permanent components in GNP and stock prices," *Journal of Economic Dynamics and Control*, Vol. 12, No. 2, pp. 255–296.

Cochrane, John H., 1997, "Where is the market going? Uncertain facts and novel theories," *Economic Perspectives,* Federal Reserve Bank of Chicago, Vol. 21, No. 6, November/December, pp. 3–37.

_____, 1994, "Permanent and transitory components of GNP and stock prices," *Quarterly Journal of Economics*, Vol. 109, February, pp. 241–266.

Daniel, Kent, David Hirshleifer, and Ananidhar Subrahmanyam, 1998, "Investor psychology and security market under- and overreactions," *Journal of Finance*, Vol. 53, No. 6, pp. 1839–1885.

DeBondt, Werner F. M., and Richard H. Thaler, 1985, "Does the stock market overreact?," *Journal of Finance,* Vol. 40, No. 3, pp. 793–808

Fama, Eugene F., 1996, "Multifactor portfolio efficiency and multifactor asset pricing," *Journal of Financial and Quantitative Analysis*, Vol. 31, No. 4, December, pp. 441–465.

Fama, Eugene F., and Kenneth R. French, 1993, "Common risk factors in the returns on stocks and bonds," *Journal of Financial Economics*, Vol. 33, No. 1, February, pp. 3–56.

_____, 1989, "Business conditions and expected returns on stocks and bonds," *Journal of Financial Economics*, Vol. 25, No. 1, pp. 23–49.

Gallant, A. Ronald, Lars Peter Hansen, and George Tauchen, 1990, "Using conditional moments of asset payoffs to infer the volatility of intertemporal marginal rates of substitution," *Journal of Econometrics*, Vol. 45, No. 1/2, pp. 141–179.

Kandel, Shmuel, and Robert F. Stambaugh, 1996, "On the predictability of stock returns: An asset allocation perspective," *Journal of Finance*, Vol. 51, No. 2, June, pp. 385–424.

Kim, Tong Suk, and Edward Omberg, 1996, "Dynamic nonmyopic portfolio behavior," *Review of Financial Studies*, Vol. 9, No. 1, Winter, pp. 141–161.

Lakonishok, Josef, Andrei Shleifer, and Robert W. Vishny, 1994, "Contrarian investment, extrapolation and risk," *Journal of Finance*, Vol. 49, No. 5, December, pp. 1541–1578.

Markowitz, H., 1952, "Portfolio selection," *Journal of Finance*, Vol. 7, No. 1, pp. 77–91.

Merton, Robert C., 1973, "An intertemporal capital asset pricing model," *Econometrica*, Vol. 41, No. 5, pp. 867–887.

_____, 1971, "Optimum consumption and portfolio rules in a continuous time model," *Journal of Economic Theory*, Vol. 3, No. 4, pp. 373–413.

_____, 1969, "Lifetime portfolio selection under uncertainty: The continuous time case," *Review of Economics and Statistics*, Vol. 51, No. 3, August, pp. 247–257.

Samuelson, Paul A., 1969, "Lifetime portfolio selection by dynamic stochastic programming," *Review of Economics and Statistics*, Vol. 51, No. 3, August, pp. 239–246.

Sargent, Thomas J., 1993, *Bounded Rationality in Macroeconomics*, Oxford: Oxford University Press.