Real theoretical advances in science do one of three things: they integrate seemingly disparate phenomena, resolve an inconsistency, or predict something new. Maxwell’s equations explained light and magnetism as being fundamentally the same, driven by photons traveling the speed of light, an integration of two seemingly different phenomena. Einstein’s special relativity theory explained the seeming inconsistency between Maxwell’s equations and Newtonian mechanics: one implying the speed of light is constant in all inertial frames, the other not. Dirac’s model of the hydrogen atom predicted the positron, which was soon discovered thereafter. The CAPM was not solving a puzzle and did not integrate two previously disparate fields; it predicted something new: that returns were related to beta and only beta, positively and linearly, through the formula

$$r_i = r_f + \beta_i (r_m - r_f)$$

Even before the empirical support was documented, in September 1971 Institutional Investor had a cover story on the CAPM, “The Beta Cult! The New Way to Measure Risk.”¹ The magazine noted that investment managers were now “tossing betas around with the abandon of Ph.D.s in statistical theory.” Unlike the theory of efficient markets, the theory of asset pricing and risk was popular from its inception.

The first tests of the CAPM were on mutual funds, presumably because they attempted to show that good mutual fund performance was a function of beta, not managerial expertise (what was soon to be called alpha). Consider the delight in some academic proving, using the mathematics in the last chapter, that a star fund manager is merely a sideshow to an incidental bet on beta. Thus, in the mid-1960s, Sharpe looked at a mere 34 funds, Jensen 115 funds, Treynor and Mazuy 57 funds.² These initial tests were not very supportive of the CAPM, but there was little consternation among the cognoscenti at that time: no “Beta Is Dead” articles would appear.
for 25 years. When scientists anticipate empirical confirmation *just around the corner*, as they did in 1970, they are really much further than they realize. Again and again, the idea is that if a bunch of smart people apply themselves to a problem that did not previously have a lot of specific attention, it is only a matter of time before it will be solved. Analogies are often made to Kennedy's exhortation to put a man on the moon, the United States's ability to increase war production in tanks and ships during World War II, or the development of a polio vaccine. Unfortunately, those are exceptions, not the rule. That the initial tests of the CAPM were not very supportive, was presumably just a data problem, though with hindsight and the now rejection of simple one-factor models, they were as doomed as a theory that purported $E = me^3$.

George Douglas published the first remarkable empirical study of the CAPM in 1969. He applied a fairly straightforward test of the CAPM. He ran a regression on 500 stocks, each with five years of data, explaining the return based on their beta, and their residual variance. With 500 different sets of betas, residual variances, and the average returns, he found that residual variance was significantly and positively correlated with returns—not beta.Lintner, one of the CAPM theoretical pioneers, also examined this, and found that residual variance and beta were both relevant in explaining cross-sectional returns.

Merton Miller and Myron Scholes (1972), two economists who would later win the Nobel Prize for independent theoretical work, pointed out several flaws in Douglas's empirical work. The important point to remember is that any empirical test is incomplete. That is, one does not test theories and reject them; in practice, testers tend to publish only tests that are consistent with their prior beliefs because they sincerely believe that to do otherwise would be a mistake. Scientists know that posterity belongs, not to the humble and virtuous, but for early discoverers of important truths. Thus, if you know the truth, as a scientist, your goal is to get published quickly proving that truth in some way so you can become one of the founding fathers.

Any empirical test makes assumptions. For example, it assumes certain variables are normally distributed, or independent, measured without error, or if they are measured with error, such errors are unbiased, or that $x$ causes $y$ as opposed to vice versa. At some level, you have to take an argument on faith. You cannot prove anything to someone sufficiently skeptical. In contrast, if the idea is less preposterous you merely ask for a few checks. If I assert that the average housing price in my city is $300,000, you may merely ask for a list of recent sales, and take this assertion on its face. If you were more skeptical, you would differentiate between stand-alone homes and townhouses, sales made in the past six months, and so on. The bottom line is that there are many legitimate issues to account for,
and generally smart people who have examined this issue have different opinions, and they can choose which evidence is important, and which is misleading or irrelevant. Any really important theory will have multiple, independent bits of relevant evidence, and advocates will emphasize that evidence consistent with their view is valid, and that evidence not consistent with their view is less valid, if not misleading or fraudulent. In the end, a scientist usually thinks herself right, not biased, and the real issue comes down to a preponderance of evidence—theoretical and empirical—sufficient for those forming the intellectual consensus among the leading researchers of the day.

Miller and Scholes found Douglas’s effect was greatly diminished by controlling for potential biases caused by

- Changing interest rates
- Changes in volatility
- Measurement errors in beta
- Correlations between residual volatility and beta
- An inadequate proxy for the market
- Cross-sectional correlation in residual errors
- Correlation between skewness and volatility

Taking these issues into account diminished but did not eliminate Douglas’s findings. It was inevitable that controlling for so many things diminishes any effect, because the more degrees of freedom you add to an explanation, the more you spread the effect among several variables as opposed to one.

Then, the seminal confirmations of the CAPM were provided by two works, Black, Jensen and Scholes (1972), and Fama and MacBeth (1973). They both applied the following “errors in variables” correction to Douglas. They formed portfolios on the basis of earlier estimations of beta, and then estimated the beta of the resulting portfolios. This would reduce the problem that plagued the earlier tests, because previously a high beta stock (for example, beta = 2) would, on average, be overestimated. That is, if betas are normally distributed with a mean of 1.0, and measured betas are simply the true betas plus error, the highest measured betas will on average have positive errors. So, a stock with a measured beta of 2, on average has a true beta significantly lower, say 1.75. On the other side, lower-than-average estimated betas would have forward betas that were closer to 1.0, so a 0.4 estimated beta would probably be closer to 1, like 0.6, in practice.

So, grouping into portfolios, and looking at the stock returns, they found the predicted result: higher beta correlated with higher returns. Even
more important, idiosyncratic volatility of these portfolios was insignificant. Please see Figure 3.1.

For both Black, Jensen, and Scholes and Fama-MacBeth, the slopes were positive—higher beta, higher return—which was all everyone interested needed to know. The only mild concern was the slope, which seemed a little flatter than anticipated. For example, in Fama and MacBeth, the hypothesis that beta is uncorrelated with returns is significant only for the entire 1935-to-1968 sample; in each of the 10-year subperiods you could not reject the hypothesis that there was no effect between beta and returns.

In spite of lukewarm support, these tests had staying incredible staying power as references that the CAPM worked. The tests were not corroborated, merely cited again and again over the next few decades. Even in the 1990s popular finance textbooks cited these papers as the primary evidence of the CAPM's validity.

For a few brief years, all was right in the world. The CAPM was slick financial theory, mathematically grounded, and seemingly empirically relevant.

But notice that the major perceived flaw in Douglas's initial work was correcting for estimation error in beta. This problem should bias both residual volatility and beta similarly, yet in Douglas's work, residual risk was significant, while beta was not. The error-in-variables correction reversed this. With the benefit of hindsight, this is totally explainable. Both beta and residual variance are positively correlated with size, a stock characteristic subsequently found highly correlated with returns. Both Black, Jensen, and Scholes and Fama-MacBeth created portfolios based on initial sorts by estimates of beta, so that the resulting spread in average beta would be maximized. But this then makes the size effect show up in the betas, because

---

**FIGURE 3.1** Returns and Betas from Black, Jensen, and Scholes (1972)
sorting by beta is, in effect, sorting by size. They could have presorted by residual variance, which is also correlated with size, and the results would have been the same as found for beta, only with residual variance the explainer of returns. Presorting by betas made the spread in betas large relative to the residual variance spread, and so made the small cap effect to speak through beta, not residual variance.

Of course, that is all with hindsight, knowledge of how the size effect works, before all the work in the 1980s showing how to correct for various biases that exaggerated the size effect. Nonetheless, it was somewhat inevitable that a theory, touted as the Next Big Thing before its first real test, is tested until the right answer is generated, and everyone stops making corrections, and cites the articles prominently for the next 20 years. As we subsequently learned, there were many more corrections to make (for example, most of the size effect was measurement error), but with the right results, there was little demand for such scrutiny. Test and make corrections until you get the right answer, then stop. In some sense, this is wisdom at work because as there are an infinite number of corrections to make, if you make them all, your results must be insignificant because the resulting sample size is so small. A skeptic may argue for adding uncertainty to the parameters, and to the parameters of the parameters, and at some point one must say no. In general, we test results with reasonable thoroughness, where reasonable is influenced by the plausibility of our result. Thus, at the heart of even the most technical debate is common sense, because one cannot relax, or validate, every assumption. But alas, common sense is no more common among academics than among anyone else.

Einstein's General Theory of Relativity was only a few years old, yet academics were eager to put the nightmare of World War I behind them, and show the common bond of the old adversaries. Proving the German's theory correct was so desired that it suggests that scientists are as objective as the rest of us. The null hypothesis, set up by standard Newtonian Physics, was that there should be a 0.85 arc-second deflection in light from stars behind the sun, while Einstein predicted a 1.7 arc-second deflection.

In 1919, an eclipse offered a chance to measure the degree to which the sun bended the rays of light from far-off stars, a prediction of the General Relativity theory. Famous English physicist Arthur Eddington was a World War I pacifist, and so had a predisposition to mend the rift

(Continued)
between German and English academics. He made a trek to the island of Principe, off the coast of West Africa, one of the best locations for observing the eclipse.

Eddington did not get a clear view of the stars because it was cloudy during most of the eclipse. He used a series of complex calculations to extract the deflection estimate from the data, and came up with an estimate of 1.6 arc-seconds. Data from two spots in Brazil were for 1.98 and 0.86, but Eddington threw out the 0.86 measurement because he was concerned that heat had affected the mirror used in the photograph, and so the standard error was too large, generating an average of near 1.7, with a relatively modest standard error.

Scientists have subsequently concluded that Eddington’s equipment was not sufficiently accurate to discriminate between the predicted effects of the rival gravitational theories, and his dismissal of the Brazilian measurement unwarranted (he could have equally applied his reservations to his own measurement). In other words, Eddington’s standard errors were too low, and the point estimate too high. He validated a theory based as much on his preconceptions as the data. Eddington died after a life full of honors, and for decades this experiment was cited as the proof of the General Theory, yet even in the 1960s, when they tried to redo the experiment given a similar eclipse and methodology, they found they could not. The tools and the event were simply too primitive to allow the kind of accuracy needed to prove general relativity.

In the late 1960s, using radio frequencies as opposed to pictures from an eclipse, Eddington’s results were, ultimately, confirmed. It is important to have correct prejudices.

Around 1980, Merrill Lynch printed large beta books, showing the beta for every stock. In 1990, William Sharpe, and Harry Markowitz won the Nobel Prize for their work in developing the MPT, which the Nobel Committee considered to be the “backbone of modern price theory for financial markets.”

**THE BEGINNING OF THE END OF CAPM**

The initial tests in the early 1970s were positive, yet more like Keynesian economics than any theory of physics: it had a consensus, but was also
empirically questionable to anyone in the know, and practitioners generally ignored it.

One of the early issues was technical in nature, having to do with sequentially estimating betas and returns. A big part of this was mere computing power. Before 1980, running a regression was difficult, and running Maximum Likelihood functions nearly impossible. These more complex functions, which are a mainstay of empirical research today, use a hill-climbing algorithm that involves a lot of steps. It is not a closed-form solution that involves a fixed set of steps as in matrix algebra in Ordinary Least Squares. So, researchers satisfied with empirical work that had issues that in hindsight are more primitive than anything one can do today in an Excel spreadsheet. Once computing got cheap, people tend to look again at early empirical work and found it could be improved by more modern techniques that were superior. For example, early work on default models used discriminate analysis because in binary modeling (default or no-default) this merely involved inverting a matrix. Logit estimation made much more reasonable technical assumptions, not requiring the errors to be both Gaussian and equal in size in both groupings. But logit requires solving a maximum likelihood, so only when the 386 PCs with their fancy built-in math coprocessors became popular was this approach feasible. Today, no one does discriminate analysis anymore except in introductory statistics classes.

A maximum likelihood approach avoided several technical problems in BJS and F&M by simultaneously estimating betas, intercepts, and slopes. Shanken (1985), Gibbons (1982), and Gibbons, Ross, and Shanken (1989) applied maximum likelihood techniques that rectified several technical issues in BJS and F&M, testing the heart of the CAPM by whether or not the market is mean-variance efficient (as implied by Tobin and Sharpe). While Shanken and Gibbons got respect for their cleverness, and rejected the CAPM at the 0.1 percent level, no one really cared about the substance of their results, because given any sufficiently powerful test, all theories are wrong. For example, Newton's ideas don't work if we measure things to the tenth decimal place, but they still work. Powerful tests on whether the CAPM was true were publishable, but everyone knew it was an almost impossibly stern standard. Such findings were published in the top journals and useful references, and probably primarily responsible for these researchers getting tenure at top research universities, but I doubt anyone was convinced as to the CAPM's practical value through these tests.

The harsh tests of Gibbons and Shanken were largely motivated by a curious result discovered by Richard Roll in 1977, when he published his famous Roll Critique. Roll proved that any mean-variance efficient portfolio would necessarily be linearly related to stock returns. It implied that to the CAPM's main implication, therefore, was merely whether or not
the market was mean-variance efficient. Roll then added, given the relevance of the present value of labor income and real estate in a person’s portfolio, the stock market was clearly not the market, clearly not mean-variance efficient, and therefore the CAPM was untestable. This might seem devastating, but Stambaugh (1982) found that inferences are not sensitive to the error in the proxy when viewed as the measure of the market portfolio, and thus while a theoretical possibility, this is not an empirical problem. Shanken (1987) found that as long as the proxy between the true market was above 70 percent, then rejecting the measured market portfolio would also imply rejecting the true market portfolio. So if you thought the market index, any one of them, was highly correlated with the true market, the CAPM was testable as a practical matter.

The Roll critique was constantly invoked to motivate the APT, in that because the CAPM necessitated the market return, and we can’t measure the market, and it could explain the zero relationship between beta and return, well, an APT approach would solve this problem. When the CAPM was being sullied in the 1990s, Roll and Ross (1994) and Kandel and Stambaugh (1996) resurrected this argument and addressed the issue of to what degree an inefficient portfolio can generate a zero beta-return correlation, which by then was accepted as fact. That is, is it possible that beta is uncorrelated with the S&P500 or whatever index is being used, even though it works perfectly with the true market index? In Roll and Ross’s words, if you mismeasure the expected return of the market index by 0.22 percent (easily 10 percent of one standard deviation away), it could imply a measured zero correlation with the market return.

This sounds devastating to tests purporting to reject the CAPM, but to generate such a null result with an almost efficient index proxy, one needs many negative correlations among assets, and lots of returns that are one hundred fold the volatility of other assets. In other words, a very efficient, but not perfectly efficient, market proxy can make beta meaningless—but only in a fairy tale world where many stocks have 100 times the volatility of other stocks, and correlations are frequently negative. Now average stock volatilities range from 10 percent to 300 percent, most between 20 percent and 60 percent annualized. Betas, prospectively, as a practical matter, are never negative.

Existence proof such as these are generally crutches for desperate promoters. It is fun to prove that there exists an equilibrium where, say, raising prices leads to more demand (a Giffen good), or where greater individual savings lowers aggregate savings (the paradox of thrift), or where protecting infant industries increases domestic productivity. These are all possible, and you can prove it rigorously given various assumptions, but they are generally untrue, because they take a lot of strange assumptions more implausible
than the main assumption that drives the basic result, and they are empirically counterfactual. Nonetheless, many people jump on these results, and using the logic similar to the idea that since I may get hit by a meteor or I may not implies that I should buy meteor insurance—suggests that existence proves these possibilities are probable. But if this is one possibility out of many, it is hardly credible. In this case, the fact that the market is not mean-variance efficient could imply a zero beta-expected return, cross-sectional relationship, but this would be highly implausible because it would imply many large negative correlations between stocks. Those who take great interest in what is possible, or impossible, are generally making unempirical statements because any real empirical issue is about probabilities, not possibilities. All the really interesting debates are about degrees of probability, not whether things are impossible or certain (in which case things are usually tautologies).\textsuperscript{13}

The big counterfactuals to the CAPM were two anomalies discovered around 1980: the size anomaly and the value anomaly. Small cap (that is, low market cap) stocks had higher returns than large stocks, while value stocks (that is, low price-to-earnings, high book value-to-market value, high dividend ratios) outperformed their opposite. Black, Jensen, and Scholes reported beta deciles with returns that had about a 12 percent annualized difference between the highest and lowest beta stocks. But for size, estimates of the difference in returns between the bottom and top size deciles was an eye-popping 15 to 24 percent.\textsuperscript{14} While the betas were positively correlated with size, that was insufficient to explain this big return differential. Something very strange was going on, and it started the hornets' nest buzzing. Within a year of the finding of the size effect, there was a special issue in the Journal of Financial Economics on this issue.\textsuperscript{15}

Over the next decade, several adjustments were made, and the size effect was drastically reduced, down from initial estimates of 15 to 24 percent to around 3 percent, annually, for the difference between the smallest and lowest capitalization stocks. Many of the biases were technical, but no less real, and affect many historical findings. The first tip-off that something was rotten in the size effect was that 50 percent of it was in January, 25 percent in the first five days of the year. What economic risk factor is so seasonal? What was to become known as the January Effect, where small stocks outperformed large stocks in the month of January, was soon found to be highly influenced by the error caused by averaging daily returns. For example, if you average daily returns, and add them up, you get a very different result from if you cumulate daily returns, and average them. If you have a very small cap stock, and its price moves from, say, its bid of 0.5 to its ask of 1.0 and back again, every day, its arithmetic return is thus +100 percent, −50 percent, +100 percent, and −50 percent, for an average daily
return of 25 percent. But in reality, its return is zero, which you would get by taking the geometric sum and dividing by the number of days. Averaging daily return for a portfolio assumes one is rebalancing a portfolio to equal-weighted starting values every day, highly impractical given one generally buys at the ask (high), and sells at the bid (lower). This arithmetic averaging issue arises quite a bit in finance. Blume and Stambaugh (1983) found this bias cut the size effect in half.16

Another big technical adjustment was discovered by Tyler Shumway, who noted that delisted stocks often have “N/A” on the month they delist, and these delisting months are actually quite devastating: down 55 percent on average!17 As small stocks delist much more frequently than large stocks, this bias overstates the return on small stock portfolios. Delisted stocks tend to overstate their actual returns because they systematically do not state these large, negative returns when they leave the database used by most researchers. The effect was almost 50 percent of the size premium. The current effect of the size effect is now around 3 percent, and even that is driven by outliers, in that if one excludes the extreme, 1 percent movers, it goes away completely.18

Size is clearly a correlated grouping, a factor, in that small cap companies are more correlated with one another than large cap companies. Many investment portfolios now make distinctions for size (micro cap, small cap, and large cap), which initially was thought to generate a big return premium, but now, it’s more like a factor people like to add to their portfolio, classic diversification, as opposed to easy money. But the return premium for the smallest cap group is only a couple of percent per year, an order of magnitude lower than what was originally discovered around 1980. The key question today is whether this factor is a true risk factor—something priced by the market, with special correlations—or whether it is survivorship bias in an unconscious research fishing expedition, or whether it’s a factor that reflects risk, or possibly a return premium from behavior biases (a “small stock aversion” effect).

In contrast, the value effect is considered stronger than the size effect, though its discovery was much more inauspicious. Initial estimates by Basu (1977) estimated that the low price-to-earnings (aka P/E) stocks outperformed the highest quintile P/E stocks by 6 percent annually, which is close to the current estimate of what the premium to a value over growth portfolio should be.19 Subsequent work played on this factor in various guises. Bhandari (1988) found that high debt-equity ratios (book value of debt over the market value of equity, a measure of leverage) are associated with returns that are too high relative to their market betas.20 Stattman (1980) documented that stocks with high book-to-market equity ratios (B/M, the ratio of the book value of a common stock to its market value) have high
average returns that are not captured by their betas. P/E, book/market, and dividend yield are all highly correlated ratios that relate a market price to some metric of fundamental value, such as book value or earnings. For some reason, the beat-up stocks, those with low earnings but even lower prices, tend to outperform. Unlike the size effect, this was not concentrated in January. Though never as popular as the size effect in the 1980s, the value effect has been more successful than the size effect over the past 30 years (size-oriented funds were started in the 1980s, while value and growth funds became popular only in the 1990s).

But one had to see size and value anomalies in context. In economics, or really any field of reporting, there is a premium on what is new, ergo, news. Anomalies get published because they are potentially interesting and useful corrections to conventional wisdom, while confirmations of existing theory are not news. But if 1,000 finance professors are trying to get published, and apply a test that, assuming the base theory is correct, still rejects the correct theory 5 percent of the time (the standard statistical metric of significant), then 50 articles are presented to journal editors with statistically significant results, and the editor is faced with either publishing noise or missing out on the next big thing. The bottom line is, there have been many ephemeral results that are published: the weekend effect, the end-of-month effect, the low-price stock effect (remember low-priced stock funds?). Even today, I find that the very anomalous Internet bubble of the late 1990s was correlated with many things: high growth rates, large R&D. Anything correlated with Internet stocks will, using the past 20 years of data, generate a huge return effect merely because of this huge bubble. One can't help but be skeptical seeing all these new anomalies, and so be more partial to theory than data. Theory is more convincing than data because data without theory are biased by its peculiarity. Knowing why is much more important than knowing what, because if you have the why correct, you can predict; the what merely describes.

Bill Schwert wrote in a 1983 review of the size effect: "I believe that the 'size effect' will join the 'weekend effect' ... as an empirical anomaly," highlighting the confusing nature of popular anomalies. The weekend effect was part of the collection of seasonal anomalies very prominent in the early 1980s. As mentioned earlier, there is the January effect, whereby small stocks outperform large stocks, the September effect (the worst month of the year), the Monday effect (worst day of the week), the Friday effect (best day of the week), and the belief that days before holidays tend to be good. In the mid-1980s, Werner De Bondt and Richard Thaler got a lot of mileage out of their documentation of mean reversion in stock prices over period of three years—this finding was essentially reversed a decade later through a much stronger momentum effect over one year, and there has been very little corroboration of the three-year reversal finding. Fischer Black remarked that
"I find theory to be far more powerful than data," after being burned in this field by false leads so many times, and Bill Sharpe noted that "I have concluded that I may never see an empirical result that will convince me that it disconfirms any theory." Understandable, but there's an unhealthy nihilism there.

The weekend effect, like the other seasonal anomalies, has disappeared. So, like the latest miracle diet or ab machine, the latest anomaly is treated skeptically by your average economist for good reason, because most are dead ends based on selection biases and just bad data.

APT TESTS

In the 1970s, before the full development of the Arbitrage Pricing Theory (APT), Barr Rosenberg (1974) suggested that portfolios demarcated by firm characteristics could likely serve as factors. Tests of the APT started in the Eighties, many pointing to the Roll critique in that the APT is inherently more susceptible to empirical validation than the CAPM, but the APT and the SDF both gave sufficient justification for doing the same thing, basically, to find some time series and use that to explain returns. If you think of the time series of stock returns, you merely need a set of equivalent time series to regress this against. But the basic approach has three distinct alternatives: the macro factor approach (oil, dollar, S&P500), the characteristic approach (size, value), and the latent factor approach (statistical constructs based on factor analysis).

In 1986, an initial test of the APT by Chen, Roll, and Ross seemed to show that the APT worked pretty well using some obvious macroeconomic factors. They used the factors representing Industrial Production, the yield between low and high risk bonds (actually, between BBB and AAA bonds, which are both pretty low risk), between short- and long-term bonds (the slope of the yield curve), unanticipated inflation, expected inflation, and finally, the market (as in the simple CAPM). These were all reasonable variables to try because they were all correlated with either the yield curve, or things investors care about. Now, given that Ross was the creator of this model, it should come as no surprise that they found these factors explained a lot of return variation. But only a few factors seemed priced, that is, you may find that stocks are highly correlated with a factor, such as the market, but on average stocks with a greater loading on this factor did not generate higher returns, the risk premium or the price of risk may be zero. This is what we mean when we say some risks are not priced (common risks that are not priced include industry risk, which clearly affects sets of stocks, though on average no industry, by itself, generates superior returns).
While this finding was promising, it was not a lot better than the CAPM in regard to its ability to explain returns, and did not explain the perplexing value or size anomalies that were confounding academics in the 1980s. More troublingly, in future tests of the APT, no one championed this particular set of factors except as a comparison, so you have people comparing an APT test using oil, employment growth, and the S&P, and another using the dollar, investment growth, and corporate spreads. As novelty is a big part of getting published, and there were no restrictions on what is allowed, the multiplicity of models examined in the literature was inevitable.

While the Chen, Roll, and Ross approach was intuitive—you could understand the factors they were suggesting—a new approach was being forged that seemed more promising: The idea that risk factors were statistical constructs, called latent factors because they were statistically implicit, not clearly identifiable by words. These are very statistical constructs that appealed to finance professors familiar with the matrix mathematics of eigenvectors and eigenvalues, a branch of mathematics that is both deep and has proven quite practical in other applications.

If we think about the original mean-variance optimization objective of investors, the latent factor approach actually makes the most sense. If one is minimizing portfolio volatility and cross-sectional volatility comes from several factors, all significant factors need to be addressed. One could come up with intuitive factors that span this space, but there are methods that are more powerful than mere intuition. Yet the first factor, which explains about 90 percent of the total factor variance, is about 99 percent correlated with the equal-weighted stock index (Connor and Koraczyk, 2008; Christopher Jones, 2001). This alone suggests that latent factors were not a good alternative. Consider that many researchers use the equal or value weighted stock index as a proxy for the market, both using eminently reasonable justification. For example, the value-weighted market proxy was used as the market by some researchers, while others used the equal-weighted market, and some used both. You can argue the value-weighted proxy is more like the market because it weights stocks by their actual dollar size, giving more weight to Microsoft than some $50 million Internet start-up from Utah. On the other hand, the equal-weighted proxy weighs small firms more, and these firms might be more representative of smaller, unlisted companies that make up the true market portfolio. The bottom line is, if either worked considerably better, we would have a story for why it should be the market proxy, and researchers would have stuck with one over the other. Neither works better than the other (their betas do not explain average returns), so it really doesn’t matter. Now, if 90 percent of the latent factor approach is indistinguishable in power to the CAPM, and there isn’t a big difference between the value- and equal-weighted index, it is improbable this approach is going to bear fruit.
But even if the latent factor approach was empirically successful, it has another large problem. Most researchers assumed there were around three to five factors, all declining rapidly in statistical relevance. But while factor 1 looked a lot like the equal-weighted index, the other factors were not intuitive at all. They were not correlated with anything you can put your finger on. Furthermore, many stocks had negative loadings on these secondary factors, so whereas the loading on the first factor, as in the CAPM, were almost all positive, these other strange factors, with no intuition, had 60 percent with positive loadings, and 40 percent with negative loadings. You could not explain exactly what the factor was, but sometimes it generated a risk premium, sometimes you were presumably paying for its insurance like properties—even though no one had enough intuition to figure out what this provided insurance against. Consider a broker suggesting that a stock is a good addition to the portfolio because it has a negative loading on the fourth factor, which no one really knows what it is, but one should be eager to earn 1 percent less because it will pay off big in some scenario when this fourth factor is really low. I am certain that this sales pitch has never happened, and considering that asset pricing theory is based on the assumption that people are in generally investing this way, is cause for skepticism.

Empirically, both of these APT approaches generated at most a modest improvement to the CAPM, and a slightly better ability to explain the perplexing size effect better. Generally, if something is true, and in the data, someone will find it, and the APT was a license to throw the kitchen sink at returns. It was rather surprising that this approach was so unfruitful.

FAMA AND FRENCH PUT A FORK IN THE CAPM

The debate changed dramatically when Fama and French published their paper in 1992. While I disagree with their diagnosis of the CAPM’s problems, their work is to be admired because it is unequaled in academic finance for its readability. Clarity and common sense are as rare among scientists as almost anyone, so when Fama and French win their Nobel Prize, as I expect they will, it will be richly deserved. Too many researchers tell you they are going to test some theory, and that it offers a rich and dynamic way to explain the post World War II data, and after reading it you have no idea what was tested or how you would use it, though the wealth of information is, indeed, rich. Fama and French don’t hide the substance of what they are doing behind abstruse mathematics; they prioritize and highlight their findings, which are often obviously relevant. When I was in graduate school, I remember my more technical professors considered Fama to be somewhat unsophisticated, good at turning a phrase (“efficient markets”) but not
doing serious research, which then were extensions of Gibbons and Shanken, in empirical work, or Banach Space extensions as suggested by the APT. I'm sure, given Fama and French's current stature, they don't remember thinking that way, in the same way few people recall being against civil rights in the 1960s.

The key finding was not so much showing value and size generate large returns in the U.S. data, it was in showing that whatever success beta had, it was completely explained by the size effect. "Our tests do not support the central production of the [CAPM], that average stock returns are positively related to beta." That is, if you remember the earlier (Figure 3.1), now we have Figure 3.2.

Shanken and Gibbons proved the CAPM wasn't perfect; Fama and French showed it was not even useful. It's not even an approximation; it's not insufficiently linear, or insufficiently positive, its point estimate has the wrong sign! Stephen Ross (1993) noted that in practice "the long-run average return on the stock, however, will not be higher or lower simply because it has a higher or lower beta." The previous single measure of expected return was accepted now as not being useful even as an incomplete measure of risk.

Such an empirical rejection would seem to be fatal, but this highlights the non-Popperian nature of real science. In Karl Popper's construct, theories produce falsifiable hypotheses. If these hypotheses are falsified, the theory is then rejected. The CAPM predicted that beta was linearly and positively related to stock returns: the actual relation was zero. End of story? It would be naïve to think that after 25 years, a simple fact would cause so many finance professors to abandon an idea that was the backbone of their

![Figure 3.2](image-url)
research papers, and formed the basis for lectures they had given for a
decade, would be snuffed out like runt on a farm. In Thomas Kuhn’s
paradigms, researchers see the seminal anomaly explained with a new theory,
and then create a new theory in the new paradigm. But the Fama-French
approach was clearly a “mend, don’t end” approach to the CAPM. The
paradigm would continue, just as the APT (or SDF), but now, a three-factor
model, the now ubiquitous three-factor F-F model

\[ r_i = r_f + \beta (r_M - r_f) + \beta_{SMB} (r_S - r_B) + \beta_{HML} (r_H - r_L) \]

where:
- \( r_m \) = market return
- \( r_S \) = small cap return
- \( r_B \) = big cap return
- \( r_H \) = high book/market return
- \( r_L \) = low book/market return

Note that this approach is really just the CAPM equation with two
extra terms. The first term, \( \beta (r_M - r_f) \), is the return on the market above the
risk-free rate, and this, plus the risk-free rate, is the CAPM. The next term,
\( \beta_{SMB} (r_S - r_B) \), is the return on the small cap portfolio minus the big cap
portfolio. As small cap firms outperform big cap companies, this average is
positive, suggesting this has a positive price, that is, it is a priced-risk factor.
The loading on this factor, \( \beta_{SMB} \), represents the sensitivity of an asset to this
factor, as measured by the beta on this factor from a regression. The final
term, \( \beta_{HML} (r_H - r_L) \), is the return on the high book to market (aka value
companies, low P/E companies) portfolio minus the return on the low book
to market (aka growth companies, high P/E companies).

This approach is totally consistent with the APT and SDF, in that either
of these motivate the addition of the value and size factors as risk factors,
because they are factors that explain a lot of cross-sectional volatility and
they are on average positive, suggesting they are priced and thus reflect
some state variable that affects our utility. Working backward, since only
risk generates average returns, the anomalies imply some kind of risk. As
the APT says, it really does not matter what or where these factors came
from, all that matters is they must be priced linearly, and so the beta terms
representing their loading tell us how much of these factors are being used.
This is known as theoretical risk factor identification, because it comes
from returns, not something more basic. The Kuhnian paradigm was not
shifted; an epicycle was added.
Saving the Standard Model

Generally, better empirical data and testing generates greater support for the correct theory over its incorrect rival. Correct theories get clearer as the data refine them. In the case of finance, the data has spawned the correlate to superstring's multiple universes (now, $10^{500}$ potential solutions), and the past is unconsciously rewritten so that a young researcher would think there has been a consistent, asymptotic trend toward the truth.

Currently, the three-factor Fama-French model remains perhaps the most popular model for benchmarking equity returns by academics, although it has a significant number of detractors. The main puzzle to the Fama-French model is what kind of risk factors do the value and size represent? Small stocks have higher betas, and are more volatile, but as these characteristics themselves are not positively correlated with returns, we can't say that small stocks outperform large stocks because of their volatility, or covariance with the market. Value stocks actually have betas that are higher in booms, and lower in busts, suggesting a win-win approach to investing that is decidedly not risky. So in what way are small size, and high-book-to-market (low P/E) stocks risky?

Early on in the size effect, people were at a loss to figure out what kind of risk that size, outside of beta, captured? Remember, the obvious risk, residual risk from these very small stocks, was diversifiable, and so not risk. Fama and French came upon the idea that both the value premium and the small stock premium were related to some sort of distress factor, that is, value stocks, whose price was beaten down by pessimists, and small stocks, which had less access to capital markets, probably had more risk of defaulting, or going bust, if the economy faltered. It may not show up in correlations or covariances, but that's merely because such risks are very episodic, like the risk of a heart attack: The first symptom of a heart attack is a heart attack.

There were two problems with this interpretation. First, as Daniel and Titman documented, it was the characteristic, rather than the factor, that generated the value and size effects. They did an ingenious study in that they took all the small stocks, and then separated them into those stocks that were correlated with the statistical size factor Fama and French constructed, and those that weren't. That is, of all the small stocks, some were merely small, and weren't correlated with the size factor of Fama-French, and the same is true for some high book-to-market stocks. Remember, in risk it is only the covariance of a stock to some factor that counts. Daniel and Titman found that the pure characteristic of being small, or having a high book-to-market ratio, was sufficient to generate the return anomaly, totally
independent of their loading on the factor proxy. In the APT or SDF, the covariance in the return with something is what makes it risky. In practice, it is the mere characteristic that generates the return lift. Fama and French shot back that their approach did work better on the early, smaller sample, and more survivorship biased 1933-to-1960 period, but that implies at best that size and value seem the essence of characteristics, not factors, over the more recent and better documented 1963-to-2000 period.31

In a similar vein, Todd Houge and Tim Loughran (2006) find mutual funds with the highest loadings on the value factor reported no return premium over the same 1975-to-2002 period, even though the value factor generated a 6.2 percent average annual return over the same period.32 Loading on the factor, per se, did not generate a return premium.

This suggests that size and value are not risk factors, just things correlated with high stock returns cross-sectionally. This could be an accident of historical data mining, as when people find that some county in some state always votes for the correct president, which is inevitable given a large sample, but ultimately meaningless. Alternatively, these findings could be proxies for overreaction. People sell small stocks too much, sell value stocks too much, and a low price leads to them to being both small cap and having a high book-to-market ratio. But this overshoot sows the seeds for an eventual correction. This effect works the opposite way for large cap and low book-to-market ratio stocks, where the prices shoot too high. Lakonishok, Shleifer, and Vishny proposed just such a model in 1994, and this is the standard behaviorist interpretation of the size and value effect: They capture systematic biases, specifically, our tendency to extrapolate trends too much to ignore the reversion to the mean.33

Another problem with the distress risk story was that, when you measure distress directly, as opposed to inferring it from size or book value, these distressed stocks have delivered anomalously low returns, patently inconsistent with value and size effects as compensation for the risk of financial distress. Around 2000, when I was working for Moody’s, I found this to be true using Moody’s proprietary ratings data. Moody’s had a unique, large historical database of credit quality going back to 1980, and I found that if you formed stock portfolios based on the credit rating of the equity’s debt, there was a perfect relationship from 1980 to 2000: the lowest-rated credits (C) had the lowest returns, followed by the next lowest-rated companies (B), then Ba, and so on, to AAA with the highest return. Better credit and lower default risk implied higher future stock returns. As not many people had Moody’s ratings data, I thought this was a very interesting finding. I presented this finding as an aside at a National Bureau of Economic Research conference, which was primarily about debt models, and I was surprised when the crowd of esteemed professors, including Andy Lo,
Robert Hodrick, and Kent Daniel, thought I had to be wrong. It is a strange thing when people you respect, and are excited to apprise of a new finding, respond by telling you that you made a mistake. They did not even have the data, but they were certain I was wrong. It was on one level frustrating, but on another encouraging, because it suggested just how valuable my fact was. That is, one thing any good idea has is novelty, and clearly this fact was novel in that no one believed it.

Like any empirical fact, other people found it as well. Ilya Dichev had documented this back in 1998, but this finding could be brushed off because he presumably had a poor default model (he used the Altman Model). But then several others documented a similar result, and finally Campbell, Hilscher, and Szilagyi (2006) find the distress factor can hardly explain the size and book or market factors; in fact, it merely creates another anomaly because the returns are significantly in the wrong direction. Distressed firms have much higher volatility, market betas, and loadings on value and small cap risk factors than stocks with a low risk of failure; furthermore, they have much worse performance in recessions. These patterns hold in all size quintiles but are particularly strong in smaller stocks. Distress was not a risk factor that generated a return premium, as suggested by theory, but rather a symptom of a high default rate, high bond and equity volatility, high bond and equity beta, and low equity return.

Finally, there is the issue of what the market is doing in the three-factor model. That is, why include the market if it does not explain equity returns? Simple, Fama and French respond. If you also include government bonds, you need the market factor to explain why the stock indexes appear to rise more than bonds do, as the factors for size and value—whatever they represent—cannot explain the aggregate difference in return between stocks and bonds. The market factor, which Fama and French admit is not relevant for distinguishing within equities, is necessary to distinguish between equities as a whole and bonds. And so it is with most models, as currency models find factors built off such proxy portfolios that help explain the returns on various currencies, but they only work in explaining currency returns. The same holds for yield curve models, where risk models are a function of those points on the yield curve that generate positive expected returns, and thus explain the yield curve by way of this risk factor. None of these factors apply outside their parochial asset class, though; they just explain their anomalies. If these were truly risk factors, and represented things that paid off in bad states of the world, it should hold across all asset classes, so that, say, a set of risk factors used to explain, say, equities, would be applicable toward currencies and the yield curve.

In the end, we have atheoretical risk factors, each of which was chosen to solve a parochial problem recursively, so that return is a function of
risk, which is a function of return, and so on. Value and size, longstanding
anomalies, were simply rechristened as risk factors, and then used to explain,
well, themselves. Fama and French said it worked out of sample by testing
it on strategies sorted by P/E ratios, but the P/E ratio is so correlated with
the B/M ratio that this is hardly out of sample. Furthermore, the inability
of the market to explain cross-sectional returns of stocks within the market,
did not hurt its importance in explaining why the market explains the
differential return between it and the bond market, even though, again,
this is tautological: The higher return of equities over bonds is explained
by the return of equities over bonds (called the equity risk premium). If it
is a risk factor, why does it apply only to itself, that is, the market, and
not to individual equities? In sum, it seems a lot more like a description
of anomalies rather than a theory, because it does not generalize to assets
it was not created from, and value and size are still factors without any
intuition.

It did not take long for researchers to jump on this bandwagon. A new
anomaly, momentum, published in 1992 by Jegadeesh and Titman, as past
winners over the last 3 to 18 months tended to continue over the next 3 to
18 months. They did not even propose a risk motivation for this finding,
and even Fama and French are reluctant to jump on that one, which would
lay bare the factor fishing strategy at work. Mark Carhart was the first to
then add momentum (a long winners-short losers portfolio) to the F-F three-
factor model and create a four-factor model in a 1996 study. At the very
least, it captured things that a naïve investor would make money on, which
was useful in Carhart’s case because in examining mutual fund returns, to
the extent a fund is relying on a simple strategy that rides on momentum,
size, or value, it seems relevant to how much pure alpha they had (although
one could argue that implementing a momentum strategy before academics
prove it exists involves alpha). That is, if you can explain a fund’s returns
through momentum, that seems important in understanding funds. Whether
it is a true risk factor is a separate issue, but the Fama-French model seems
useful in explaining it irrespective of whether its factors are related to risk,
or are predictive.

SERIAL CHANGES TO APT

The striking fact about multifactor models applied to cross-sectional equities
is that there is no consensus on the factors. The APT and SDF approaches
have slightly different emphasis in empirical tests, and the APT testers gen-
erally have more intuitive results, whereas the SDF approach generates more
statistically powerful, but less compelling results, but generally it all comes
down to identifying factors and using them to explain returns. Which returns? Well, after Fama and French, everyone tried to explain, which really means to correlate, a model with the returns of portfolios that are created by cross-tabbing book-to-market and size, two characteristics that are historically related to returns for some reason. That is, each month, you sort stocks into quintiles by size, and then sort within each quintile by book or market. You get 25 portfolios after the 5 by 5 sort. Many of the proposed models seem to do a good job explaining these portfolio returns as well as the three-factor Fama-French model, but it is an embarrassment of riches. Reviewing the literature, there are many solutions to the Fama-French set of size-value return, but there is little convergence on this issue. This is especially true given the great variety of factor models that seem to work, many of which have very little in common with each other.

For example, Chen, Roll, and Ross (1986) assert six factors as the market return, inflation (expected and actual), industrial production, Baa-Aaa spread, and yield curve innovations; Sharpe (1992) suggests a 12-factor model, while Connor and Korajczyk (1993) argue there are "one to six" factors in U.S. equity markets using principal components analysis. Ravi Jagannathan and Zhenyu Wang (1993) assert human capital and the market portfolio are the two factors because a metric of human capital is needed to capture the total market; they later argue (1996) that time-varying betas can explain much of the failure in CAPM. Lettau and Ludvigson (2000) use the consumption-wealth ratio in a vector autoregressive model to explain cross-sectional returns. Jacobs and Kevin Wang (2001) argue idiosyncratic consumption risk and aggregate consumption risk; and Jagannathan and Yong Wang (2007) have year-over-year fourth quarter consumption growth. This research is vibrant and ongoing, but the diverse approaches extant suggest they have not begun to converge, even within the prolific Jagannathan/Wang community. The bottom line is that unlike the value, momentum, or size, no one has created an online index of these more abstract factors because they change so frequently; there is no single such factor that would be of interest to general researchers.

The problem is that 25 portfolios sorted by factors that were found to be correlated with returns makes for an easy target for statistical explanation. In data mining, you throw a bunch of time series against a bunch of data until a high correlation is found, and then, poof, an explanation. It's derisively known as "survival of the fittest," meaning the set of explanatory data with the highest R^2 or statistical fit is the answer, and given enough data to throw at these poor 25 portfolios, one set of these will have an R^2 this will be rather high. The key point to remember is the hot alternative in 1994 looks nothing like the alternative in 2000, which looks nothing like the current alternative. These solutions invariably explain the 25 Fama-French portfolios sorted on
value and size, and then do horribly on different tests, and the best evidence of this is the ephemeral nature of these solutions; a correct solution would draw emulators, but mere statistical correlates are one-shot publications. Fama and French created a standard for getting published that is good for just that, getting a publication, and then, extensions are null results, uninteresting, and so everyone moves on.

Another issue is that with financial time series, many patterns are cyclical but not predictable. That is, if you look at the past time series for the value and size factor proxies, you will see that if you could merely predict their major swoons and booms, you will be able to explain a lot of the spread between size and book- or market-sorted portfolios. Given that these turning points are so important, and there are only a handful of them, what other time series have similar turning points? Any such series would be found to explain stock returns. Small cap stocks did poorly from 1969 through 1974, rebounded through 1983, trailed through 2000, but then rebounded smartly. Value stocks did poorly in the Internet bubble, but also rebounded nicely. With thousands of economic and financial time series, some combination will generate similar turning points merely by chance. Thus, you look at consumption growth, and if that doesn’t work, take real consumption growth, then real consumption growth per capita, and if that doesn’t work, use year-over-year (as opposed to quarterly seasonally adjusted) consumption growth, and then one can deduct expenditures on durable goods, or use only durable goods. The U.S. Department of Commerce’s National Income and Product Accounts shows hundreds of variations on expenditures and consumables; all you need is one them to turn on the correct dates, which one eventually will, and all are, on first glance, plausible proxies for the SDF. The key is when these valuable proxies are found sequentially, they lose all their statistical properties because they are not random samples. Yet, this bias is often not conscious but merely the result of a process that publishes only significant results, and so these biases are often totally unintentional. Please see Figure 3.3.

In Figure 3.3, a roughly noisy set of data are presented in dots. A linear best fit line has only two degrees of freedom, an intercept and slope, whereas the curvy line is a fifth-order polynomial and has six degrees of freedom. Although the polynomial function passes through each data point, and the line passes through none, the line is a better fit because if the function that created these dots was extrapolated, one should expect the polynomial model that exactly fit the data would do much worse out of sample. Overfitting is related to the wisdom of Occam’s razor, which says simpler is better. We all want to explain the data better, but there’s a trade-off where at some point it hurts our ability to predict. This is very nonintuitive because people generally find that more information always adds to one’s confidence in a
subject. That is, most people would consider 13 reasons for some proposition is better than 7 reasons. However, statistically, a model with 13 parameters often does worse at prediction, largely because the correlation between the regressors causes the standard error of the forecast to blow up. For example, in a simple model with two explanatory bits of data, $x_1$ and $x_2$, the standard errors are proportional to $1/(1 - \text{corr}(x_1, x_2))$, so the higher the correlation with existing information, the higher the standard errors for the individual coefficients, which implies a more precarious out-of-sample performance. This generalizes to more variables in a straightforward way.

Thus, the number of turning points for a model explains the data, which is about 10. A couple of hundred really smart Ph.D.s do research, each trying to explain the 25 Fama-French portfolios, and a couple do. Seems great! Each individual researcher is probably not biased, but the collective process is. The problem is that no one knows the number of degrees of freedom in a model, because that is implicit in a process of an unknown number of unpublished approaches.

**SKEWNESS**

Recent research on distributional adjustments has been intriguing, but not compelling. For example, Harvey and Siddique (2000) found that a firm’s marginal contribution to negative skew, coskewness, can do as well as the Fama-French three-factor model. Yet, the data seem fragile. They calculate skew using 60 months of data for each stock, and this biases the data set because it excludes about 40 percent of the sample. Ang, Chen, and
Xing (2002) find that downside beta, which is like negative coskewness, is positively related to returns for all but the highly volatile stocks, where the correlation with returns is negative. Furthermore, value and small companies that have high returns, have positive skewness. Thus, skewness may be at work, but as a general measure of risk, it is quite fragile.

One major problem with this research is that, prospectively, all stocks have positive skew. A stock with negative skew is rare, if only because for any stock, the upside is unlimited, but they can all go to only zero. Furthermore, stocks with the greatest downside have the greatest upside, so you cannot simply say the risk of failure is a sign of skew: these companies invariably have the greatest chance of both massive success and total failure. Thus, in looking for a preference for skew within equities, this is like evaluating the claim that height explains basketball prowess among NBA players. Clearly, being tall is helpful in being a good basketball player, yet, conditional upon being in the NBA, the effect is weak at best.

Interestingly, individual stocks are on average positively skewed, while portfolio returns are negatively skewed because correlations between stocks increase in market downturns. Thus, if anything, the risk is greater for a portfolio of stocks by this measure than for individual stocks, and we do see some evidence of a risk premium for stocks in aggregate, as opposed to cross-sectionally.

The real world is much more complicated than any model, but models are about compression. Pattern prediction is strangely seductive, as a method for generating a pattern like something is often mistaken as a specific prediction. But my ability to generate a time series that looks like a price series from the New York Stock Exchange from a glance at a graph and matching mean, variance and skew, does not mean this is valuable. The key is, can we identify some of these parameters as more or less risky, and do they correlate with future average returns? No, which is why theories like chaos are not used 40 years after their discovery.

Yet despite the long history of looking at higher moments of variable distributions, such as skew and kurtosis, its anemic results and therefore relatively small empirical examination can make it appear unexamined. I have been in meetings many times where someone proposes ignoring upside variance as a measure of risk, as it seems intuitive and rarely used, so I would venture hundreds of times every day, worldwide, people in finance independently generate this hypothesis. But a really high return tends to give information relevant to whether there will be a large drawdown, because 100 percent returns suggest a high probability of a 50 percent drawdown, just as much as a previous 50 percent drawdown does. Thus, as a forward looking metric of risk, the semi-deviation generally does worse than metrics that use all the information. The bottom line is that adjustments to the
normal distribution were there at the beginning, presented as a solution every so often (see Kraus and Litzenberger (1976), and Post and van Vliet (2004)), and if it worked, these would be a canonical non-normal model by now.48

Initial theoretical work by Rubinstein (1973) showed how one could add skewness and kurtosis with the CAPM.49 The idea is that sometimes people prefer positive skew, as in preferences for lottery tickets, and for negative skew, as when people appear to overinvest in investments with large but infrequent losses. The Journal of Portfolio Management online list of abstracts mentions skewness in no fewer than 66 articles and kurtosis in 44. Why do these intuitive corrections seem to always be talked about, but never make it to the status of, say, the Fama-French value or size factors? Because the results are so weak and nuanced.

**ANALOGY TO BUSINESS CYCLE FORECASTING**

The empirical arc, from optimism that we had a useful model of asset return, to pessimism that we do not, has happened before in economics. The intuition is that real data follow patterns like the tides, the motion of the planets, or the distribution of fauna along a riverbank. If you take something big and important, like the stock market, and say, we have finally generated comprehensive data on it, many smart people will assume there is a pattern there just waiting to be discovered.

So it is no coincidence that the initial development of the CAPM was the early 1960s, which coincided with the creation of the Center for Research in Security Prices (CRSP). At the behest of Merrill Lynch, two professors at the University of Chicago, James H. Lorie and Lawrence Fisher, created what was to become the preeminent database on stocks in the United States. In 1964, their database was complete, and they successfully demonstrated the capabilities of computers by analyzing total return—dividends received as well as changes in capital as a result of price changes—of all common stocks listed on the NYSE from January 30, 1926, to the present. A seminal article by Lorie and Fisher in the Journal of Business reported the results. The article proclaimed that the average of the rates of return on common stocks listed on the NYSE was 9 percent. The front page of the New York Times financial section heralded the pair’s findings.50

Researchers had data, just in time to merge with the new theory. For a researcher used to picking apart dog-eared books where many have treaded before, this truly was exciting. In a statement that can only be understood by financial quants, Rex Sinquefeld, now CEO of the quantitative equity management firm Dimensional Fund Advisors, noted, “If I had to rank
events, I would say this one (the original CRSP Master File) is probably slightly more significant than the creation of the universe.\textsuperscript{51} Virgin data, important data, and theory, imply one is like Kepler looking at Brache’s data on the positions of the planets—new laws are about to be confirmed, or discovered. The initial researchers would be \textit{Founding Fathers} of this new field of scientific finance!

Alas, not everything moves in patterns like the planets. A similar life cycle of untempered optimism, adding epicycles to explain anomalies, and then manifest failure, could be applied to macroeconomics and its study of business cycles. In the 1930s, you had Simon Kuznets and Jan Tinbergen creating vast accounting systems for the entire nation, what was to become Gross National Product and National Income and Product Accounts. As Keynes was developing the first macro theory at that time that explained the dynamics of a macroeconomy from such building blocks as investment, consumption, and so on, it was thought that one would finally see everything in front of them and know how to steer. As Keynes wrote in 1940 contemplating the arrival of national income accounting data, “We are in a new era of joy through statistics.”\textsuperscript{52} The idea was that if we can see where we are and where we have been, we know where we are going. Keynesian theory would allow us to adjust the economy, so that unlike before when we would sail, hit an iceberg, sail again, repeat, it would simply be smooth sailing.

Jan Tinbergen won the first economic Nobel prize for this work, and over the next 10 years several Nobel prizes were bestowed on researchers in this area, because it was important and seemed obvious that given our ability to measure, and the application of sophisticated mathematics, the ability to manage was \textit{just around the corner}. Paul Samuelson’s first paper in 1939 was to apply mathematics to the new theory of macroeconomic dynamics, in this case a second-order difference equation, and the optimism was palpable.\textsuperscript{53} Macroeconomic experts were about to wander in the desert for 40 years.

But a generation of failure is not apparent in real time, only with the benefit of hindsight a generation later. By the early 1970s, macroeconomic modeling had missed many economic turning points, but still many were jumping into the field and saw a bright future in macroeconomic forecasting. After all, it is easy to explain away prior failures, and as business cycles happen only once every 5 to 10 years, it was not like these failures were happening all the time. There was the oft-repeated joke “He predicted seven of the last four recessions,” in full confidence that it was just a matter of time. By the late 1970s, macroeconomics was being assaulted by an unanticipated increase in interest rates, and the unanticipated simultaneously high levels of unemployment and inflation, and leading edge researchers were seeing a crisis. The models were breaking down, becoming obviously
wrong to the casual user. But practical reputation follows achievement with a lag. In the 1980s, I worked with economists who worked for the Bank of America in the mid-1970s, and they talked of a whole floor of economists, forecasting at various industry and regional levels (Modesto, California, Retail Sales employment growth, for example). The chief economist then was very charismatic, and had a conspicuous trophy wife he took to corporate functions.

Right out of college, having been a teaching assistant for the macroeconomist Hyman Minsky, I was excited to be an economist myself. I got a job at First Interstate, a major bank in Los Angeles, in the economics department, making very little money doing very mundane analyst work, but seemingly learning from priests who knew what was true and important. I could not have been happier. The first thing I discovered was that we had no ability to forecast the future, and our explanations of the past were lame. No one had a clue why inflation was falling so fast at that time, the main financial trend of 1986. No one really understood why oil prices were falling so low (then down to $13 from $30), why the stock market crashed in 1987, or whether the net effect on the United States would be positive (low inflation) or negative (wealth effects on spending). We had lots of forecasts for all sorts of variables, often to the second decimal place, and followed the rule to “forecast early and often,” so that by the time data arrive, the reader is comparing them to the latest revision, which given the inertia in time series, allows one to get close with the most recent forecast. The key is to overload your reader with data so they will have misplaced the forecast from last year, or worse, three years ago, and just look at the latest paper when the actual number comes out. Nothing imprints as strongly as learning as a young man, something you strongly believed turned out mistaken.

It became clear to me we were not fooling anyone because economics departments were shrinking. Nothing speaks louder than declining employment in a field. My enthusiasm for macroeconomics was clearly not shared by market participants. When I got back into banking after graduate school around 1994, the large regional bank I worked for had over 10,000 employees and 1 economist, whose main job was public relations, not advising internal decision making, and this was a typical use for an economist. A few years later, they got rid of him.

In the macroeconomic equivalent to the Fama-French paper, in 1980, Chris Simms showed that a simple Vector Autoregression (VAR) could do as well as a hundred equation macro model. That is, instead of modeling the economy as a complex set of subroutines that operate in a giant feedback loop based on optimizing behavior of firms and consumers, just regress past changes in GDP on past changes in GDP and investment growth. Such an atheoretical approach worked just as well (or badly) in terms of predicting
the future as the more theoretical models derived from first principles, and
embodying the monetarist and Keynesian pet theories. When such a simple
approach dominates a much more complicated approach, there is always
soul searching.

But hope always springs anew. In the late 1980s, one of my North­
western professors, Mark Watson, tried to create a better leading economic
indicator. The Leading Economic Indicators was created in the 1950s,
before computers, and a wealth of econometric methods. It seemed obvious
that the top professors in the field could easily improve upon this indica­
tor, because fundamentally there was no "rational expectation" reason why
forecasting recessions, over the next six months, should be so intrinsically
difficult. That is, you can say that no matter how much information and
processing power, the stock market is hard to predict because every day
we predict the next day and this involves predicting how we will predict
the next day, so we are predicting predictions. To the extent something is
predictable, it gets into prices now, not in some foreseeable lump in the
future. But that argument doesn't apply so much to business cycles, which
are based, not on expectations, but real activity. So it seemed that predicting
GDP growth was theoretically possible, and from a pure technology and
technique perspective, an obvious trade-up from the naïve rules created in
the 1950s that became the Leading Economic Indicators (LEI).

The standard LEI averaged 10 indicators thought predictive of the econ­
omy. Other than normalizing each indicator to make its volatility equivalent,
they were simply added together in the late 1960s. In contrast, Stock and
Watson developed a fancy Kalman filter approach and applied to state spaces
in a Markov to create a new and improved LEI in 1989. This different ap­
proach seemed to make the Sims critique irrelevant, because those macro
modelers were using too much theory, and not enough statistics. After all,
the Kalman filter was actually being used to guide rockets, and, as we all
know, rocket scientists are the gold standard of modeling stochastic pro­
cesses. They subsequently went on a whirlwind tour of major central banks
with their state-of-the-art forecasting device. The LEI was a simple sum of
interesting variables that was created in 1950, the Kalman filter was cutting
edge econometrics, and tested on data from 1919 to 1985, dominated the
naïve LEI in backtesting. It was like the difference between my cell phone
and two cans with a string attached to their ends. Or so it seemed.

The problem is that there is much less data there than you think, be­
cause the data are cyclical. With cyclical data, if you can call the peaks and
troughs, you explain a lot of the data. So really, there were only eight reces­
sions after World War II, and God knows how many iterations of the Stock
and Watson model before development. As any good economist who knows
the data really well, the data used in her model become in sample rather
unconsciously, in that anyone looking at historical time series of macroeconomics for 10 years has seen how various indexes relate to one another, and as the data change modestly each year, the data to be explained are pretty much static—until the next recession.

Unfortunately, the real-time performance of the Stock and Watson approach has been no better than the original, naïve Leading Economic Indicators index. In 1990, the difference was that the dollar was strengthening and the yield curve was steepening, meaning that monetary policy was not particularly tight immediately before the recession, in contrast to the prior recession. In 1993, the then head of the American Economic Association remarked at the annual meeting, that regarding the 1991 recession, “established models are unhelpful in understanding this recession.” The Stock-Watson model failed to predict the 1990–1991 recession, and an updated version of the model (one that would have caught the 1990 recession) then failed to predict the 2001 recession. Stock and Watson (2003) discuss, with admirable honesty and clarity, this failure and argue that it is hard to predict recessions because each is caused by a unique set of factors. For instance, housing and durable goods consumption was strong preceding and throughout the 2001 recession, and the decline was focused on high technology manufacturing. By contrast, in the 1990–1991 recessions, housing and durable goods spending slowed considerably. As Stock and Watson say, “Without knowing these shocks in advance, it is unclear how a forecaster would have decided in 1999 which of the many promising leading indicators would perform well over the next few years and which would not.” Indeed.

All recessions have key differences in their origins and emphasis, because businessmen and politicians tend not to repeat the exact same mistakes, other than that a large minority of them will have overinvested in something that, with hindsight, was bound to fail. Thus, in 1990, the key losers were hotels, in 2000, telecoms, and in 2008, the primary losers were mortgage lenders. There is little relation to the overbuilding and overpricing in hotels, telecoms, and housing other than they subsequently contracted, and so were at abnormal peaks when viewed with hindsight at the peak of the expansion.

When a field has new data and theory, there is a natural belief that some basic laws will be discovered or confirmed. It is as if one expects there to be higher level patterns from a bunch of lower level decisions, because we know that most patterns, at a lower level, cannot be seen. The idea of stepping outside one’s little universe, like the two-dimensional man in Flatland seeing his universe from the third dimension, is incredibly alluring, seems like you should see more. Further, advances in statistics and computing imply that patterns previously unseen should now be revealed. Alas, sometimes a lot of data just highlight that we do not have any good theories.
SUMMARY

The CAPM started with a tepid confirmation, and while there was much whistling past the graveyard, and hopes that warts like the size and value effect would go away, it turned out they merely focused researchers on questions they should have asked at the very beginning. The seeming correlation between beta and returns was mainly due to the correlation with the size effect, and this, in turn, was mainly due to measurement errors. The best cross-sectional predictor of equity returns is momentum and book-to-market, neither of which have an intuitive risk rationale. Tweaks to the model have been conspicuous in their ephemeral nature, always popping up in different guises.

"In the end," notes Fischer Black, "A theory is accepted not because it is confirmed by conventional empirical tests, but because researchers persuade one another that the theory is correct and relevant."58 This is a very postmodern interpretation of finance, accurate in the short run, but not in the long run.