Michael Cooper
Purdue University

Huseyin Gulen
Virginia Tech

Is Time-Series-Based Predictability Evident in Real Time?*

We show that out-of-sample tests used in the time-series predictability literature may suffer from test size problems related to the common practice of exogenous specification of critical parameters, such as the choice of predictive variables, traded assets, and in-sample estimation periods. We perform specification searches across these parameters and find that rejections of the null hypothesis of no predictability are very sensitive to minor variations in parameter specification. We perform simulations to determine if the observed predictability in the data is real. The simulations suggest that much of the literature's out-of-sample evidence of time-series-based predictability is consistent with data snooping.

There now appears to be overwhelming evidence of stock market predictability. A large body of research shows that excess returns on the aggregate market are forecastable from the default spread, dividend yield, dividend payout, the term spread, consumption data, inflation, industrial production, wealth, and labor income, to name but a few variables.[1] Yet, despite this seemingly overwhelming evidence, there appear to be few real-world investors capable of taking advantage of this time-series predictability, especially at the levels of profits suggested by the academic pre-

* We thank Mike Cliff, David Denis, Diane Denis, John Easterwood, Ken French, Roberto Gutierrez Jr., Yufeng Han, Greg Kadlec, Raman Kumar, John McConnell, Raghu Rau, Vijay Singal, an anonymous referee, and Albert Madansky (the editor); seminar participants at the University of Georgia, Purdue, and the 2002 Western Finance Association meetings for their helpful comments and suggestions; and James Garret for his excellent research assistant work. Cooper gratefully acknowledges financial support from an International Research Award from the Center for International Business Education and Research. Also, we thank Allan Timmerman, Peter Bossaerts, and Sydney Ludvigson for providing us with the data used in their studies. Contact the corresponding author, Huseyin Gulen at hgulen@vt.edu.

[1] A partial list of academic papers that document stock market predictability include Keim and Stambaugh (1986), Campbell (1987), Campbell and Shiller (1988a, 1988b), Fama and French (1988, 1989), Breen, Glosten, and Jagannathan (1989), Cochrane (1991), Ferson and Harvey (1991), Hodrick (1992), Lamont (1998), Pontiff and Schall (1998), Lewellen (1999), Lettau and Ludvigson (2001), and Santos and Veronesi (2006).

[*Journal of Business*, 2006, vol. 79, no. 3]

0021-9398/2006/7903-0009$10.00

dictability papers.[2] As Cochrane (1999) states, “It is uncomfortable to note that fund returns still cluster around the (buy-and-hold) market Sharpe ratio” (68). He suggests that “if the strategy is real and implementable, one must argue that funds simply failed to follow it” (68). Thus there appears to be a large gap between real-time investor performance and the high levels of predictability found in the literature.

We offer an explanation for this performance gap that is based on potential collective data-snooping biases on the part of researchers. This collective snooping may be inherent to the market predictability literature because (1) there is little explicit guidance from theory regarding the identity of the predictive variables used in these studies, hence making it a data-fitting exercise; (2) any new research endeavor is inherently conditioned on the collective knowledge built up to that point; and (3) there is a tendency in the literature and the profession at large to retain the findings that “work” and discard the ones that do not. Given these issues, it is feasible that a nontrivial proportion of the relations reported in the literature, and accepted as economically meaningful, are simply due to pure luck. As Denton (1985), Lo and MacKinlay (1990), Black (1993a, 1993b), Foster, Smith, and Whaley (1997), Sullivan, Timmermann, and White (1999), Conrad, Cooper, and Kaul (2002), and Ferson, Sarkissian, and Simin (2003) point out, we (usually out of sheer necessity) *collectively* condition our studies on existing empirical regularities with the unintended consequence of snooping the data. In this paper, we attempt to gauge the impact of potential data snooping on empirical findings in the return predictability literature that are based on commonly used methodologies and under plausible scenarios of snooping.

In the market predictability literature, researchers typically use “out-of-sample” tests. In these tests, researchers use a rolling (recursive) forecast method in which a model of expected returns is estimated using prior in-sample data. Parameter estimates from in-sample periods are used to create forecasts in holdout periods. Out-of-sample tests have traditionally been used to avoid model overfitting. The tests are guaranteed to work, however, only asymptotically. In finite samples, overfitting and, hence, spurious evidence of predictability may emerge when a large number of models are considered. The goal of this paper is to demonstrate that such is the case in return predictability studies.

Because of data limitations, most out-of-sample studies are not truly out-of-sample, in the sense of using an independent holdout period. Typically, researchers use the same, or substantially the same, period to discover pre-

[2] The current notion that the stock market is predictable stands in contrast to the well-documented inability of mutual funds to beat the market (see Carhart 1997; Wermers 2000). It is interesting to note that in addition to mutual fund studies, nearly all other studies of real-time investment performances also fail to show that the market is clearly beatable. Barber and Odean (2000) find this for individual investors; Christopherson, Ferson, and Glassman (1998) for pension funds; Pirinsky (2001) for banks, investment advisors, and insurance companies; Desai and Jain (1995) for “superstar” money managers; Metrick (1999) for newsletter recommendations; and Barber et al. (2001) for analysts’ consensus recommendations.

dictive relations as to test them. If snooping occurs, then the use of full-period information can result in a subtle but important test size problem emanating from a researcher's design of the out-of-sample forecasting algorithm. The problem is related to the degree of endogeneity (or lack thereof) used by researchers in constructing such tests. Specifically, many features of a researcher's out-of-sample experiment such as the choice of assets, predictive variables, length of the in-sample window used to obtain forecast parameters, and model selection methods are typically *exogenously* determined by the researcher after having obtained familiarity with the entire data.[3]

As a first pass to examine the potential effects of snooping, we construct a universe of out-of-sample models by considering plausible combinations of a limited set of forecast parameters. Specifically, using a recursive forecasting method that is ubiquitous to the market forecasting literature and using returns and predictive variables from recent studies, we perform an exhaustive specification search over a range of three parameters that are typically exogenously specified: predictive variables, assets, and in-sample window lengths (henceforth referred to as "econometrician choice variables").[4] For expositional purposes, we refer to the results of these specification searches as "exogenous out-of-sample forecasts." This process of explicitly snooping the data yields some interesting insights. First, our results show that many models do appear to "beat the market." However, second, the successful models span the range of the choice variables, providing us with little guidance on the true values of these parameters. Third, rejections of the null hypothesis of no predictability are very sensitive to minor changes in the number and identity of predictive variables, to changes in the in-sample window length, and to the criterion

[3] Some researchers have examined endogenizing (i.e., allowing the data to choose in an ex ante manner) one or two of these forecasting aspects in the time-series predictability literature. Pesaran and Timmermann (1995) and Bossaerts and Hillion (1999) endogenize predictive variable selection by using various statistical model selection criteria, including a new method, predictive least squares–Markov dimension criteria (PLC-MDC); Pesaran and Timmermann (1995) develop a "hyper-selectivity" forecasting model that endogenizes the statistical model selection criteria; and Pesaran and Timmermann (2002), hoping to solve issues related to model nonstationarity by capturing shifts in factor/return relations, endogenize in-sample window length. Swanson and White (1997) endogenize variable selection and window length via linear models and artificial neural networks in an attempt to forecast macroeconomic variables. In the cross-sectional literature, Cooper, Gutierrez, and Marcum (2005) further explore such "real-time" issues inherent in out-of-sample tests by requiring the investor to endogenously determine in-sample the optimal predictor variables, rules relating those variables to future returns, and the sort dimensionality. Once they endogenize these portfolio investment parameters, it is difficult for an investor to outperform a passive buy-and-hold benchmark portfolio.

We thank the referee for pointing out that exogenous determination of model sets is not a problem per se; it does not induce a bias when out-of-sample tests are performed on *new* data. However, in studies of market predictability, it is often the case that predictive models are discovered and tested on the same data.

[4] It is worth emphasizing that these three aspects appear to us to be the most obvious exogenously specified parameters. There are many other more subtle parameters a researcher must specify before implementing an out-of-sample test. They include return horizon, model selection criteria, asset allocation rules, forecast update frequency, test(s) of the null, learning features, transaction costs, and others. In a later section of the paper, we expand our analysis to endogenize model selection criteria and transaction costs.

used to evaluate the null. Fourth, the best model specifications tend to be similar to the models reported in the literature. Overall, the potential for serious data-snooping problems is high, especially considering that there appears to be minimal ex ante guidance from theory on the identity of the true models.

We then attempt to gauge the effects of our collective snooping on the best specifications from the exogenous out-of-sample forecasts. We devise a measure of data snooping that measures the proportion of "real" profits observed in the exogenous out-of-sample forecasts that can be generated using "random variables" to predict actual returns as we iterate over plausible specification searches. Our simulations show that inadvertent snooping biases, perhaps based on our collective familiarity of the data, have the potential to explain most, if not all, of the observed predictability in the best real-data models. Specifically, depending on the grouping of predictive variables and assets and the particular test statistic used, we find that after just one specification search, the amount of predictability from the best random-data specification is approximately 40%–96% of what we observe in the real data. This is particularly striking given that the best specification from the real data is likely to be upward biased, since it comes from our ex post knowledge. After just 15–20 iterations, we observe that the best random-variable specification is in the range of 65% to greater than 100% of the predictability observed in the real data. This suggests that only 20 researchers, each performing a specification search across predictive variable combinations and in-sample window lengths, could generate predictability equal to that found in the real data, even though the random variables have no real predictive ability.

We also examine variants of commonly employed recursive tests to examine whether the best models from the real-data specification searches are obtainable in "real time." We implement out-of-sample experiments in which we remove the effects of parameter snooping by endogenizing the econometrician choice variables in a recursive framework, as in Pesaran and Timmermann (1995). We find that the decrease in predictability between the best exogenous out-of-sample forecasts and the real-time, endogenized forecasts is quite large, with two out of the three data sets we examine showing no evidence of predictability, and one data set showing some marginal ability to "beat" the market, but only in a zero transaction costs setting.

The remainder of the paper is organized as follows. In Section I, we develop and discuss a "reality spectrum" of out-of-sample forecasts, guided by the underlying principle that the correct approach to an out-of-sample forecast should be to simulate as accurately as possible all the uncertainties faced by a real investor. The reality spectrum provides a summary of many parameters in the market forecasting literature that are typically exogenously specified. We use this reality spectrum as the basis for the design of the exogenous out-of-sample forecasts in Section II. In Section III, we present the results of simulations and out-of-sample forecasts that endogenize the selection of predictive variables, assets, in-sample window lengths, and model selection criteria. Section IV contains our conclusion.

I. A Reality Spectrum of Out-of-Sample Forecasts

We first develop an idea of the extent and identity of exogenous parameter specification in the market predictability literature. To facilitate this, table 1 provides a "reality spectrum" of out-of-sample forecasts. The left-most column of the table lists of some of the more commonly exogenously specified choice variables. We separate the choice variables into "major" and "minor" categories on the basis of our perception of their relative importance in the market predictability literature. The reality spectrum ranges from "none," in which a researcher employs an in-sample methodology, up to "high," in which all the choice variables are endogenized. Clearly, even a high level of realism in the modeling process is a simplified version of an actual investors' decision-making process.[5]

The choice of predictive variables is likely to be the most commonly exogenously specified parameter, with many papers invoking the phrase "we focus on a common set of lagged instruments, shown to have worked in previous studies" as justification for their chosen set of predictive variables. Studies typically use a fixed set of three to five variables. For example, Campbell (1987) uses lagged returns, T-bill yield, change in yield, and a yield spread measure. Keim and Stambaugh (1986) use the yield on Baa-rated bonds less the one-month T-bill yield, a ratio of the level of the S&P 500 to the 45-year average of the S&P 500 level, and a measure of share price, averaged equally across the quintile of smallest market cap firms on the New York Stock Exchange. Other prominent variables in the literature include dividend yields (Shiller 1984; Fama and French 1988), dividend payout, or the ratio of dividends to earnings (Lamont 1998), term spread measures (Fama and French 1989), and the level of consumption relative to income and wealth (Lettau and Ludvigson 2001). Nontraditional variables include deseasonalized cloud cover, raininess, and snowiness (Hirshleifer and Shumway 2001), ambient noise level (Coval and Shumway 2001), and the distance of a trader from the corporate headquarters of the traded stock (Hau 2001). A casual perusal of DataStream (a popular purveyor of worldwide financial data) reveals thousands of time series for the United States and many other countries. Thus, as Foster et al. (1997) point out, "There are limitless possible linear and nonlinear transformations of these variables" (593).

The next most commonly varied parameter is likely to be the predicted asset(s). Popular assets include excess U.S. stock (equal-weighted and value-

[5] It is likely that any modeling attempt cannot possibly account for the myriad of uncertainties facing a real-time investor. For example, a more accurate depiction of the "real-world" uncertainties facing an investor might include a real-time expanding predictive variable set (likely numbering in the tens or hundreds of thousands of variables), a survivorship bias–free collection of assets within all countries and across all countries (Brown, Goetzmann, and Ross 1995; Goetzmann and Jorion 1999), and a real-time consideration of all possible model selection methods and computing technologies, to name just a few. Therefore, our conservative depiction of the number of econometrician choice variables serves to likely bias our tests in favor of finding predictability.

TABLE 1 A Reality Spectrum of Out-of-Sample Forecasts

	Reality Level				
Econometrician Choice Variables	None (In-Sample)	Low (Out-of-Sample)	Some (Out-of-Sample)	More (Out-of-Sample)	High (Out-of-Sample)
Major:					
Predictive variables	Exogenous	Exogenous	Endogenous	Endogenous	Endogenous
	Fixed	Fixed	Fixed	Fixed	Real time expanding
Assets	Exogenous	Exogenous	Exogenous	Endogenous	Endogenous
	Single or multiple	Single	Multiple	Multiple	Real time expanding, including failed assets
In-sample window lengths	Exogenous	Exogenous	Exogenous	Endogenous	Endogenous
	Single	Single	Multiple	Multiple	Many
Minor:					
Model selection	Exogenous	Exogenous	Endogenous	Endogenous	Endogenous
	None	None	Single criterion	Competing criteria	Many competing criteria
Trading rule	Exogenous	Exogenous	Exogenous	Endogenous	Endogenous
	Single rule	Single rule	Single rule	Multiple rules	Many rules
Return horizon	Exogenous	Exogenous	Exogenous	Exogenous	Endogenous
	Single	Single	Single	Multiple	Multiple
Forecast update frequency	Exogenous	Exogenous	Exogenous	Exogenous	Endogenous
	In-sample method	Single holdout period	Single holdout period	Recursive	Recursive
Study period	Exogenous	Exogenous	Exogenous	Endogenous	Endogenous
	Fixed	Fixed	Subperiods	Subperiods	Many subperiods
Test of null	Single test	Single test	Single test	Multiple tests	Multiple tests
Learning	No	No	No	Yes	Yes
Transaction costs	No	No	Exogenous	Endogenous	Endogenous
			Fixed	Fixed	Time-varying
Technology	Exogenous	Exogenous	Exogenous	Endogenous	Endogenous
	Fixed	Fixed	Multiple	Multiple	Real-time expanding

NOTE.—The left-most column lists commonly specified econometrician choice variables. We separate the choice variable into “major” and “minor” categories on the basis of our perception of their relative importance in the market predictability literature. The reality spectrum ranges from “none,” in which a researcher employs an in-sample methodology and exogenously determines the choice variables, up to “high,” in which the choice variables are endogenously determined by the data.

weighted indexes from the Center for Research in Security Prices and the S&P 500 composite index) and bond portfolios (Keim and Stambaugh 1986). Also popular are industry-grouped portfolios (Ferson and Harvey 1991), size-sorted portfolios (Lo and MacKinlay 1997), and size- and book-to-market-sorted portfolios (Ferson and Harvey 1999). Internationally, some papers have used the Morgan Stanley Capital International (MSCI) indexes (Ang and Bekaert 2006). As with the predictive variable list, the above list of assets is by no means all-inclusive.

The next econometrician choice variable, in-sample window length, might be less of an obvious parameter as compared to predictive variables and asset choice. But as we will show in the next section, it has a large bearing on how often one rejects the null of no predictability. In-sample window length refers to the sample period from which model parameters (betas) are estimated. These betas are then multiplied by the predictive variable realizations to form expected return estimates in step-ahead out-of-sample periods. The choice of window length is not at all straightforward; if one believes that regime shifts may have occurred across a given sample period, one may employ a relatively short fixed-length window or apply exponentially declining weights to past observations. If one believes that "the truth" emerges only from betas estimated over a long time series, one may employ an expanding or long fixed-length window. Examples of studies that use expanding windows include Pesaran and Timmermann (1995), Ferson and Harvey (1999), and Lettau and Ludvigson (2001). Examples of studies that use fixed windows include Bossaerts and Hillion (1999), Sullivan et al. (1999), and Ferson and Harvey (1993, 1999).

A goal of our paper is to evaluate potential data-snooping effects in the market predictability literature arising from a researcher's freedom in selecting the above "major" choice variables, with major being defined in large part from the range in values these variables exhibit in the literature. However, there are many other "minor" parameters that a researcher must calibrate in implementing out-of-sample forecasts, and in the interest of presenting a more complete reality spectrum, we discuss these next. We begin with model selection. This category is closely related to "choice of predictive variables." Obviously, a researcher studying time-series predictability must choose a group of conditioning variables. For papers that exogenously choose and hold fixed their predictive variables (which appear to us to be the majority of papers in the market predictability literature), this category is effectively removed, or is what we will refer to as "none." However, there are papers that endogenize the choice of predictive variables from within an exogenously specified fixed universe of variables. For example, Pesaran and Timmermann (1995) endogenize model selection across a family of statistical and economic-based model selection criteria (e.g., R^2, Akaike, Schwarz, "sign," "Sharpe," and "wealth" criteria). Bossaerts and Hillion (1999) endogenize variable selection across a number of exogenously specified statistical selection methods, including PLC-MDC. Swanson and White (1997), Brown, Goetzmann, and Kumar (1998), and Allen and Karjalainen (1999) use various forms of nonlinear selection criteria (including neural networks and genetic algorithms) to endogenize variable selection.

Finally, in table 1 we list other less obvious parameters that a researcher must decide on in implementing an out-of-sample forecast. They include the trading rule used to translate expected return forecasts into portfolio weights, return horizon of the predicted assets (potential values include monthly, quarterly, yearly, etc.), forecast update frequency (e.g., monthly, yearly), study period (typically researchers will use data up to the point of their study, with the starting point being the exogenously determined aspect), test(s) of the null hypothesis (e.g., parametric or nonparametric test statistics, parameterization of a utility function, method of standard error calculation, and the number and identity of "risk factors" in estimating an alpha), forms of learning, and transaction costs. Finally, the appropriate use of "technology" is an important issue that is rarely addressed in studies of predictability.[6] For example, it would be inappropriate to use a computer-intensive genetic algorithm to uncover evidence of predictability before the algorithm or computer was available.

II. Mimicking the Process of Time-Series Predictability Research

Does the inherent process of research tend to make us converge on values of the econometrician choice variables that work the best, but are not known ex ante, in real time? In this section, we explicitly search over combinations of choice variables that researchers may have explored and examine the robustness of the resulting models.

A. Predictive Variables and Assets

Table 2 describes the data. We use data from three recent market predictability papers: Pesaran and Timmermann (1995), Bossaerts and Hillion (1999), and Lettau and Ludvigson (2001).[7] The data sets from these papers include time-series variables such as a consumption to wealth ratio, dividend yield, dividend payout ratio, various interest rate and term structure measures, a default risk measure, inflation, industrial production, a January dummy, and a number of other predictive variables, along with the excess returns of 13 countries' major indexes (in U.S. dollar returns) covering 1953–98. The Lettau and Ludvigson data use quarterly excess returns, and the other two data sets use monthly excess returns. In the literature, each one of the variables from the three data

[6] An exception is Pesaran and Timmermann (1995). In their real-time recursive study of predictability on the S&P 500, they consider a subgroup of forecasting tools that "use simple statistical and computing techniques that were clearly publicly available to any investor throughout the sample period analyzed in this paper" (1203).

[7] We thank Allan Timmerman, Peter Bossaerts, and Sydney Ludvigson for providing us with the data used in their studies.

TABLE 2 Data

		Data Period			
		In-Sample			
Data Set	Index	Initial Period	Final Period	Out-of-Sample	Variables
1	S&P 500	1953(9)–1973(6)	1953(9)–1997(12)	1973(9)–1998(3)	CAY, SPX, DY, DP, RREL, TRM, DEF
2	S&P 500	1954(1)–1963(12)	1954(1)–1992(11)	1964(1)–1992(12)	DY, EP, Tbill$_{-1}$, Tbill$_{-2}$, Tbond$_{-1}$, Tbond$_{-2}$, Π, ΔIP, ΔM
3	Australia	1971(4)–1981(3)	1971(4)–1995(4)	1981(4)–1995(5)	JAN, $R_{i,-1}$, $R_{i,-2}$, $R_{Bi,-1}$, $R_{Bi,-2}$, YTM$_{\mathrm{B}i,}$, P_i, Tbill$_i$, DY$_i$, PE$_i$
	Belgium	1981(3)–1991(2)	1981(3)–1995(4)	1991(3)–1995(5)	
	Canada	1980(1)–1979(12)	1970(1)–1995(4)	1980(1)–1995(5)	
	France	1979(3)–1989(2)	1979(3)–1995(4)	1989(3)–1995(5)	
	Germany	1970(4)–1980(3)	1979(4)–1995(4)	1980(4)–1995(5)	
	Italy	1973(4)–1983(3)	1973(4)–1995(4)	1983(4)–1995(5)	
	Japan	1981(4)–1991(3)	1981(4)–1995(4)	1991(4)–1995(5)	
	Netherlands	1971(4)–1981(3)	1971(4)–1995(4)	1981(4)–1995(5)	
	Spain	1978(2)–1988(1)	1978(2)–1995(4)	1988(2)–1995(5)	
	Sweden	1982(4)–1992(3)	1982(4)–1995(4)	1992(4)–1995(5)	
	Switzerland	1980(1)–1989(12)	1980(1)–1995(4)	1990(1)–1995(5)	
	United Kingdom	1970(9)–1989(8)	1970(9)–1995(4)	1980(9)–1995(5)	
	United States	1970(1)–1979(12)	1970(1)–1995(4)	1980(1)–1995(5)	

NOTE.—The table provides a summary of the data used in this paper. Data set 1 includes quarterly excess returns of the S&P 500 and seven predictors: estimated trend deviation in consumption (CAY); S&P excess return lagged once (SPX); dividend yield lagged once (DY); dividend payout ratio lagged once (DP); relative T-bill rate, calculated as the 30-day T-bill rate minus its 12-month moving average lagged once (RREL); the term spread (10-year T-bond yield less one-year T-bond yield) lagged once (TRM); and the default spread, calculated as the yield difference between BAA and AAA corporate bonds lagged once (DEF). Data set 2 includes monthly excess returns of the S&P 500 and nine predictors: DY; S&P 500 aggregate earnings-to-price ratio lagged once (EP); one-month T-bill rate lagged once (Tbill$_{-1}$) and twice (Tbill$_{-2}$); 12-month T-bond rate lagged once (Tbond$_{-1}$) and twice (Tbond$_{-2}$); yearly inflation rate lagged twice (Π); change in industrial production lagged twice (ΔIP); and change in narrow money stock lagged twice (ΔM). Data set 3 includes monthly excess returns, in $US terms, of 13 countries' indices and 10 predictors specific to each country. The 13 indices are the S&P 500 (United States) and the country indices reported by MSCI for the remaining 12 countries. The 10 predictors are a January dummy (JAN), monthly stock return of the local index in $US terms lagged once ($R_{i,-1}$), monthly stock return of the local index in $US terms lagged twice ($R_{i,-2}$), monthly bond excess return lagged once ($R_{Bi,-1}$), monthly bond excess return lagged twice ($R_{Bi,-2}$) yield to maturity on a representative Treasury bond lagged once (YTM), price level of the country's market index lagged once (P_i), the yield-to-maturity on a three-month Treasury bill lagged once (Tbill$_i$), the stock market's dividend yield lagged once (DY$_i$), and the stock market's price-to-earnings ratio lagged once (PE$_i$). For each data set, we report the initial in-sample, final in-sample, and out-of-sample periods.

sets has been shown to predict returns. Thus our data set provides us with a comprehensive and plausible set of predictive variables to carry out our tests.

B. Exogenous Out-of-Sample Forecasts

We construct out-of-sample forecasts using all possible models formed by combinations of predictive variables, assets, and commonly employed in-sample window lengths within each of the three data sets. We want to stress that we do not believe that any one researcher actually conducted such a search, but that the process of research, across researchers and over time, may have implicitly resulted in such a search. We follow these steps for each data set to construct the forecasts.

1. For all possible variable combinations, I ($I = 2^K - 1$ models [each model includes an intercept], where K is the number of predictive variables in each data set), and in-sample window lengths, W ($W =$ 10, 15, and 20 years and an expanding window for data set 1 and $W = 5$ and 10 years and an expanding window for data sets 2 and 3),[8] and all possible assets, A ($A = 1$ for data sets 1 and 2, and $A = 13$ for data set 3), we construct an out-of-sample time series of returns using the following recursive approach:

   A. We estimate, using ordinary least squares (OLS), a linear model of the form $r_\tau = \beta_I' X_{\tau-1,I} + \varepsilon_{\tau,I}$, where $X_{\tau-1,I}$ is an $(n_I + 1) \times 1$ vector of predictive variables, including a vector of ones for the intercept term, and r_τ is the excess return for asset A during in-sample period τ. We estimate the model during the in-sample period W and use the in-sample loadings on the predictive variables to form expected return forecasts in recursive, step-ahead, out-of-sample periods. For example, consider data set 2. The initial in-sample period is 1954(1)–1963(12). We estimate the linear model, obtain predictive variable loadings, and form an expected return estimate in the first out-of-sample period in 1964(1).

   B. We then roll forward the in-sample end date by one period, reestimate the model, and obtain a forecast for 1964(2). We repeat this process until the end of the out-of-sample period. Thus, for each data set, we obtain $W \times A \times (2^K - 1)$ out-of-sample forecast series.

2. For each out-of-sample forecast series, we obtain a series of realized returns from the following trading strategy: go long asset A if the expected excess return estimate for that period is greater than zero, else invest in a T-bill. For each return series we estimate four performance measures: a forecast beta, Jensen's alpha, the Fama-French

[8] We use longer in-sample window lengths for data set 1 because of its quarterly return horizon.

three-factor model alpha, and the Henriksson and Merton (1981) market-timing measure.[9]

We present the results in table 3: panel A for data set 1, panel B for data set 2, and panel C for data set 3. In each panel, we report the percentage of models that reject the null hypothesis of no predictability at the 5% level or better for each of the four performance measures. We also break down the rejection rates by in-sample window length *W* and the number of predictive variables *K* in a given model. Considering all exogenous combinations of the three econometric choice variables of assets, predictive variables, and in-sample window lengths results in 508 combination for data set 1 (1 asset × 4 windows × $[2^7 - 1]$ models), 1,533 combinations for data set 2 (1 asset × 3 windows × $[2^9 - 1]$ models), and 39,897 combinations (13 assets × 3 windows × $[2^{10} - 1]$ models) for data set 3. The large number of exogenous forecast combinations might seem extreme, but we maintain that when it is considered in light of our reality spectrum of table 1 and in terms of the observed variations of these parameters in the published literature, it is not, but rather likely represents a *smaller* number of combinations relative to the true number of specifications from which the best-performing models in the literature have been drawn.

In table 3, there are large variations in predictability across variable groups, in-sample window lengths, data sets, and performance measures. Depending on which performance measure is used, we find evidence of out-of-sample predictability in approximately 8%–22% of the exogenous combinations for data set 1, 53%–78% of the exogenous combinations for data set 2, and 1%–5% of the exogenous combinations for data set 3.[10] Obviously, this is a huge variation, and it illustrates the striking differences in predictability across exogenously specified variable groups, assets, in-sample window lengths, and performance measures. The level of predictability in the best-performing models is high: in data set 1 (as reported in the bottom of table 3) the best model (CAY and RREL, with a 10-year window), as defined by terminal wealth (terminal wealth is the total wealth at the end of the out-of-sample period from investing one dollar at the beginning), handily beats an S&P 500 buy-and-hold benchmark ($40.03 vs. $18.99), has a quarterly Jensen's alpha of 1.14% ($p = 0.004$), a Fama-French three-factor alpha of 1.1% ($p = 0.009$), a forecast beta of 0.68 ($p = 0.02$), and a market-timing value of 1.15 ($p = 0.04$). We observe similar performance for the best model combinations in

[9] We thank Kenneth French for providing us with the monthly premiums for the Fama-French three-factor model (Fama and French 1993). The forecast beta (β_f) provides a measure of overall out-of-sample fit and is calculated by regressing the monthly realized return of the predicted asset on the forecasted return from each forecasting model: $r_\tau = \alpha + \beta_f r_{\text{forecast},\tau} + \varepsilon_\tau$.

[10] As Fama (1991) points out, all tests of asset pricing models are conditional on the model of risk adjustment used, and the results in this section dramatically demonstrate different rejection rates across commonly employed test statistics. We do not directly pursue this issue, but obviously the exogenous choice of test statistic could dramatically change conclusions of predictability for these three data sets.

TABLE 3 **Rejection Rates for Out-of-Sample Forecasts Formed from Specification Searches across Exogenously Specified Predictor Variables and In-Sample Window Length Combinations**

		Percentage of Out-of-Sample Forecasts Rejecting the Null under the Following Criteria:			
Prespecified Variable	Number of Specifications	Forecast Beta $\beta_f > 0$ ($p_{\beta f} \leq .05$)	Jensen's α $\alpha_j > \alpha_{j,bh}$ ($p_\alpha \leq .05$)	FF α $\alpha_{\text{ff}} > \alpha_{\text{ff},bh}$ ($p_{\alpha ff} \leq .05$)	Market Timing $\text{HM}_{p1+p2} > 1$ ($\text{HM}_p \leq .05$)
		A. Data Set 1[a]			
1 variable	28	28.6	25.0	25.0	32.1
2 variables	84	27.4	21.4	19.0	19.0
3 variables	140	28.6	13.6	10.7	6.4
4 variables	140	28.6	7.9	7.1	3.6
5 variables	84	19.0	6.0	3.6	.0
6 variables	28	3.6	3.6	3.6	.0
7 variables	4	.0	.0	.0	.0
10 years	127	14.2	7.1	6.3	8.7
15 years	127	26.0	8.7	8.7	9.4
20 years	127	40.9	14.2	13.4	10.2
Expanding	127	19.7	18.1	12.6	2.4
Total	508	21.5	11.4	10.0	8.4
		B. Data Set 2[b]			
1 variable	27	11.1	18.5	37.0	70.4
2 variables	108	32.4	47.2	50.0	77.8
3 variables	252	51.6	60.7	59.9	79.8
4 variables	378	64.8	61.6	61.1	81.2
5 variables	378	72.2	61.1	57.9	80.2
6 variables	252	77.4	62.7	56.7	82.1
7 variables	108	79.6	56.5	49.1	81.5
8 variables	27	77.8	59.3	40.7	77.8
9 variables	3	66.7	33.3	33.3	66.7
5 years	511	87.7	37.2	35.0	51.9
10 years	511	40.3	51.9	47.6	92.2
Expanding	511	65.8	88.8	88.3	97.1
Total	1,533	60.6	53.2	51.4	78.2
		C. Data Set 3[c]			
1 variable	390	0	2.1	1.8	2.8
2 variables	1,755	.4	2.9	2.7	3.9
3 variables	4,680	.9	2.8	2.3	4.2
4 variables	8,190	1.2	2.4	2.2	4.2
5 variables	9,828	1.2	2.2	2.2	4.3
6 variables	8,190	1.1	2.5	2.4	4.7
7 variables	4,680	.9	2.2	2.3	4.9
8 variables	1,755	.8	2.2	2.5	5.8
9 variables	390	1.3	2.1	2.1	7.4
10 variables	39	2.6	2.6	0	7.7
5 years	13,299	.1	2.2	2.8	4.8

TABLE 3 (*Continued*)

		Percentage of Out-of-Sample Forecasts Rejecting the Null under the Following Criteria:			
Prespecified Variable	Number of Specifications	Forecast Beta $\beta_f > 0$ ($p_{\beta f} \leq .05$)	Jensen's α $\alpha_j > \alpha_{j,bh}$ ($p_\alpha \leq .05$)	FF α $\alpha_{\mathrm{ff}} > \alpha_{\mathrm{ff},bh}$ ($p_{\alpha ff} \leq .05$)	Market Timing $\mathrm{HM}_{p1+p2} > 1$ ($\mathrm{HM}_p \leq .05$)
10 years	13,299	1.5	2.7	2.1.	4.9
Expanding	13,299	1.4	2.3	2.0	3.7
Total	39,897	1.0	2.4	2.1	4.9

NOTE.—This table presents the percentage of out-of-sample forecasts rejecting the null hypothesis of no predictability under various performance measures. The out-of-sample forecasts are formed from all exogenous combinations of predictive variables and in-sample window lengths for three data sets and are based on a recursive methodology. For each out-of-sample forecast combination, we obtain a series of realized returns from the following trading strategy: go long in the traded asset if the expected excess return estimate is great than zero, else invest in a t-bill. The performance measures are the forecast beta (β_f), Jensen's alpha, the Fama-French (1993) three-factor alpha (FF alpha), and the market timing statistics of Henriksson and Merton (1981) (HM_p and HM $_{p1+p2}$). Rejection rates are reported at a 5% or better significance level. The rejection rates reported for the Jensen's alpha and FF alpha are based on two conditions: the alpha of the out-of-sample portfolio must be greater than the alpha of the buy-and-hold portfolio and the alpha of the out-of-sample portfolio must be significant at the 5% or better level. The coefficient estimate of the slope (β_f) provides a measure of overall out-of-sample fit and is calculated by regressing the monthly realized return on the forecasted return $r_\tau = \alpha + \beta_f r_{\mathrm{forecast},\tau} + \varepsilon_\tau$. For β_f we report the percentage of forecasts with positive betas and significance betas at the 5% or better level. In the last row, we report the total number of model combinations and the average rejection rates across variable and window combinations.

[a] We examine all $\binom{7}{K}$ forecasting model combinations, where $K = 1, 2, \ldots, 7$, of the seven predictive variables and four window lengths of 10 years, 15 years, 20 years, and expanding, for a total of $4 \times (2^7 - 1) = 508$ out-of-sample return series from 1973(9)–1998(3). The traded asset is the S&P500, using quarterly returns. Model with highest terminal wealth (TW): variables: CAY, RREL; window: 10 years; TW: $40.03; $\mathrm{TW}_{\mathrm{bh}}$: $18.99. Model with lowest TW: variables: DY, DEF; window: 10 years; TW: $5.68; $\mathrm{TW}_{\mathrm{bh}}$: $18.99.

[b] We examine all $\binom{9}{K}$ forecasting model combinations, where $K = 1, 2, \ldots, 9$, of the nine predictive variables and three window lengths of 5 years, 10 years, and expanding, for a total of $3 \times (2^9 - 1) = 1{,}533$ out-of-sample return series from 1964(1)–1992(12). The traded asset is the S&P500, using monthly returns. Model with highest TW: variables: EP, Tbill_{-1}, Tbond_{-1}, Tbond_{-2}, Π, ΔIP, ΔM; window: expanding; TW: $84.25; $\mathrm{TW}_{\mathrm{bh}}$: $17.55. Model with lowest TW: variables: EP, IP, ΔM; window: 10 years; TW: $5.58; $\mathrm{TW}_{\mathrm{bh}}$: $17.55.

[c] We examine all $\binom{10}{K}$ forecasting model combinations, where $K = 1, 2, \ldots, 10$, of the 10 predictive variables and three window lengths of 5 years, 10 years, and expanding, for 13 countries, for a total of $13 \times 3 \times (2^{10} - 1) = 39{,}897$ (or 3,069 for each country) out-of-sample return series from 1980(1)–1995(5). The traded assets are the S&P 500 monthly returns for the United States and the $US denominated MSCI monthly returns for each of the 12 other countries. Model with highest ratio of TW relative to $\mathrm{TW}_{\mathrm{bh}}$: Italy; variables: $R_{i,-1}$, $R_{i,-2}$, Tbill_i; window: 5 years; TW: $9.32; $\mathrm{TW}_{\mathrm{bh}}$: $2.01. Model with lowest ratio of TW relative to $\mathrm{TW}_{\mathrm{bh}}$: Australia; variables: JAN, $R_{i,-1}$, $R_{i,-2}$, $R_{Bi,-1}$, $R_{Bi,-2}$, YTM_{Bi}, DY_i; window: 5 years; TW: $0.85; $\mathrm{TW}_{\mathrm{bh}}$: $3.75.

data set 2. In data set 3, there is some combination of variable group and window that results in market beating performance in every country. The annualized spread in the Fama-French alphas between the best and worst model combination is approximately 9.4%, 8.7%, and 17.8% for data sets 1, 2, and 3, respectively (not reported in the tables).[11] Thus, on an ex post basis, there is a large degree of out-of-sample predictability evident in all three data sets.

It is not too surprising that certain model specifications generate out-of-sample predictability when we consider that the predictive variables in each data set were selected from other successful papers that were to some extent

[11] The spread for data set 3 is the average spread across countries. The highest (lowest) annual spread for the best model from an individual country occurs in Sweden (the Netherlands) at 34% (11%).

contemporaneous to the studies from which we gathered our data. What is more surprising (and troublesome) is that, first, the distributions of the econometrician choice variables in the successful forecasts span the full spectrum of the choice variables and, second, rejections of the null are very sensitive to minor changes in the choice variables. Thus the specification searches do not offer us much guidance on the true values of these econometrician choice variables.

For the first point above, consider the number of variables in a model, as broken out in each panel of table 3. Across the data sets and performance measures, there is no consistent pattern: in some cases larger variable groups result in more rejections of the null (e.g., in panels B and C, we observe more rejections across the four performance measures as we move from models with one variable up to models in the five- to 10-variable range), but in other cases we observe greater numbers of rejections for smaller-variable models (e.g., in panel A, we observe a greater rate of rejection of the null for models with one to five variables and then a sharp drop-off for models with six and seven variables). We also calculate, but do not report in the tables, the inclusion rates of the predictive variables in the successful and unsuccessful specifications. Some variables, such as CAY, lagged market, and the dividend payout ratio, have similar inclusion rates across the successful and unsuccessful models. Others, such as dividend yield and default spread, show up more frequently in the unsuccessful models but are also included in a nontrivial number of successful models.

Consider next the in-sample window length. In data set 1, from panel A of table 3, the 20-year window is best for the forecast beta criterion, the three-factor alpha, and the market-timing measure, and the expanding window is best for Jensen’s alpha. In data set 2, a five-year window is best for the forecast beta measure, but an expanding window is best for the other three measures. Finally, in data set 3, an expanding window is never the best: the window that results in the most rejections is often of intermediate length. However, for all data sets, we still observe rejections of the null for all window lengths.

For the second point, that rejections of the null are sensitive to minor changes in the econometrician choice variables, consider the best model from data set 1 as reported in note a of table 3. The best-performing out-of-sample model as defined by terminal wealth is CAY and RREL using a 10-year in-sample window. If we vary the window to 15 years, the terminal wealth drops from $40.03 to $30.43 (not reported in the tables).[12] If we add DY to the CAY and RREL model, we see a maximum wealth of $28.55 for a 20-year window and a minimum wealth of $13.81 for a 15-year window. For data set 2, as reported in note b of table 3, the best model, which uses an expanding window, has a terminal wealth of $84.25. If we keep the same set of predictive

[12] The composition of the models in the top and bottom deciles of terminal wealth is available from the authors on request.

variables but change a five-year window, the wealth drops to \$22.02. The same types of patterns are evident for the international data, in panel C. Indeed, small changes to the predictive variable group and in-sample window can result in dramatic changes to our inferences of predictability.

This section illustrates that it is possible to find a great degree of predictability, or none at all, in all three data sets by looping over relatively small ranges of just three econometrician choice variables. Overall, the lack of dominant predictive variables and estimation periods, the sensitivity of the forecasts to minor changes in parameterization, and the general lack of theory to guide us on the correct specifications suggest at best that, as a group, the successful models emanate from "richly complex processes" or, at worst, arise from some combination of luck and/or ex post snooping.

C. The Use of the Best In-Sample Model in "Out-of-Sample" Tests

If we consider that *all* the variables in the three data sets have been used and continue to be used in the related literature, then the potential for data-snooping problems may be large in light of the above evidence. As we mention in the introduction, there are multiple ways in which (inadvertent) snooping may occur. One simple method is the widespread practice of identifying the best model(s) from a series of in-sample tests and then testing that model "out-of-sample" using the same, or substantially the same, data. In this subsection, we examine how models identified on the basis of R^2 perform in out-of-sample tests conducted on the same period in which the highest-R^2 model was selected. If the use of the best in-sample model in same-period out-of-sample tests results in a bias toward rejecting the null, we would expect to observe that the best in-sample models are among the best specifications in the exogenous out-of-sample forecasts.

We estimate full-period (in-sample) regressions for each data set. For data set 1, the highest-R^2 model, across all window lengths, is always in the top 7% of the exogenous out-of-sample models (on the basis of Jensen's alpha). For data set 2, the best in-sample model, using an expanding window, is ranked second out of 1,533 runs on Jensen's alpha. Interestingly, when the best model is run out-of-sample for the other window lengths, it is at the fiftieth percentile or below for Jensen's alpha. For Japan, in data set 3, the best in-sample model, using an expanding window, ranks in the top 20%, at 533rd out of 3,069 runs. For the other 12 countries, the results are similar. These results suggest that using variables that have "worked" over the entire sample period will bias the recursive out-of-sample performances of these variables toward providing evidence of predictability. Our results in this subsection are consistent with those of Bossaerts and Hillion (1999), who show that choosing the best model in-sample and then testing whether it works using parameter estimates that rely on the past only induces a strong bias toward the appearance of predictability.

III. Is It Real?

A. Simulations

In this section, we attempt to provide the reader with a sense of the number of specification searches required, using random data as predictive variables, to generate the same level of predictability as we observe in the real data. We devise a measure of data snooping that measures the proportion of "real" profits observed in the exogenous out-of-sample forecasts that can be generated using random variables and iterating over all possible specification searches for each of the three data sets. In each iteration of our simulations, we generate a set of random variables that by design have no correlation with the dependent variable. Specifically, for each data set in table 2, we use nonrepeating seeds to generate random $N(0, 1)$ predictive variables. For each data set, we generate K random variables, where K is the number of predictive variables in each data set. For each data set, we then generate out-of-sample forecasts using all possible specifications of the random variables combined with the in-sample window lengths. The random variables have no time-series relations with the market return data and therefore should, on average, have no out-of-sample predictive ability. We run the simulation 100 times and report how the best random-data models compare to the best real-data models.

Figure 1*a* reports the results for data set 1, 1*b* for data set 2, and 1*c* for data set 3. The *x*-axis reports the number of simulations, and the *y*-axis reports the percentage of the best real-data performance obtained by the best random variable model for five measures: terminal wealth, a forecast beta, Jensen's alpha, the Fama-French three-factor model alpha, and the Henriksson and Merton (1981) market-timing measure.

If the reader believes that the entire profession has conducted just *one* complete specification search of data set 1, (i.e., has examined just once the $4 \times [2^7 - 1] = 508$ models generated from combinations of seven variables and four window lengths), then figure 1*a* suggests that between 56% and 93% of the predictability in the real data is consistent with a snooping bias, depending on which test statistic the researcher chooses to consider. Perhaps more shocking is that at 10 iterations of the simulations, the random-data models obtain 100% or greater of the real-data predictability for all five performance measures. For the forecast beta and market-timing measures, this hurdle is exceeded after just three and two iterations, respectively. After approximately 40 runs, the level of predictability from the best random-factor model as a percentage of the best real-data model asymptotes to between 110% (for the three-factor model alpha) and approximately 180% (for the forecast beta).

The results for data sets 2 and 3 also suggest that evidence of predictability using those assets and predictive variables is to a large degree consistent with snooping. In data set 2, which uses monthly returns of the S&P 500 and nine common predictive variables, we see that after just one run, snooping across variable groups and window combinations generates between 39% (forecast beta) and 92% (market timing) of the corresponding measures from the real data. After 15 runs, the number increases to 65%–96%. For data set 3, we report the simulation results for France, since France represents an average country in terms of real-data predictability (as judged by terminal wealth of the best real-data model relative to the terminal wealth of a buy-and-hold strategy).[13] After one random-data specification search for France, we see that the best random-data model generates between 40% (forecast beta) and 95% (market timing) of the real-data predictability. After 15 runs, this increases to 100% (forecast beta) to 111% (Jensen's alpha). For all three data sets, the level of predictability in the real data that is explained by the random-data simulations is high. However, for data set 2, this level is not as high as with the other data sets, suggesting that some of the predictability in the best real-data specification (perhaps 3%–35%—the amount varies across test statistics) may not be consistent with snooping.

We also examine the predictability of the top 20 and top half of all specifications (based on each of the five performance measures) from the random data as a percentage of the corresponding number of best models from real data (not reported in the figures). For all three data sets, we observe that it takes a relatively small number of specification snoops to find predictability that approaches the levels found in the real data. For example, for data set 1, only 14 iterations are required before the average of the top 20 simulations obtains profits equal to or greater than those found in the real data. This suggests that snooping problems in typical time-series models are not solely restricted to the best model, but could also affect other related or even non-related models (i.e., robustness tests) that arise from a study's pool of candidate predictive variables. This is troublesome because it implies that robustness tests, in which a researcher adds or deletes a variable from a successful predictive model, could easily be generated that give the appearance of supporting a particular model but are in fact consistent with snooping.

B. Endogenizing the Choice Variables via a Recursive Strategy

We view the simulation results as quite striking, but they do not prove that time-series predictability in the real data is false—only that the results are *consistent* with snooping. Thus, to address the question of whether the performance of the best models from the real-data specification searches is obtainable in "real time," we implement out-of-sample experiments in which we attempt to remove the effects of parameter snooping via endogenizing the selection of the econometrician choice variables in a recursive framework, similarly to Pesaran and Timmermann (1995) and Cooper et al. (2005). This method employs an in-sample period to choose the best combination of predictive variables and in-sample window lengths and then uses the optimal

[13] Results of the simulations for the other 12 countries are consistent with those for France and are available on request.

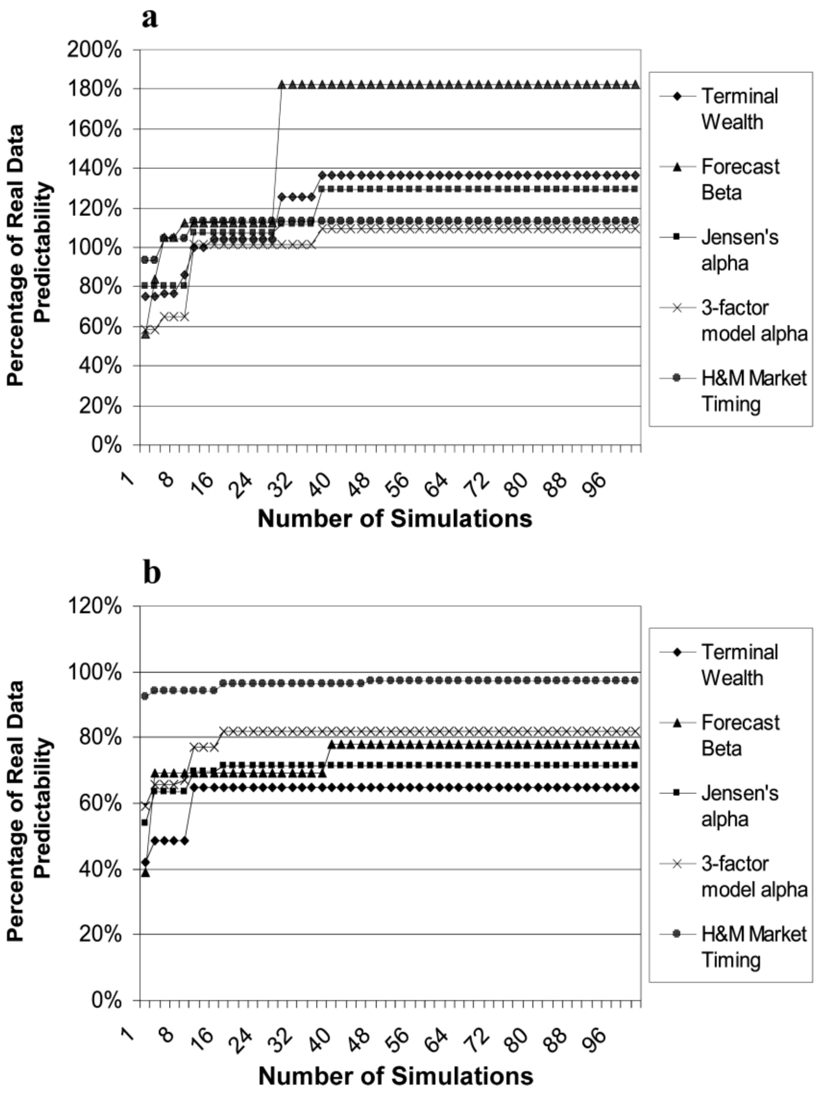

FIG. 1.—The number of specification searches required (*x*-axis), using random data as predictive variables, for the best random-data model to generate a given percentage of the predictability from the best real-data model (*y*-axis) for terminal wealth, a forecast beta, Jensen's alpha, the Fama-French three-factor model alpha, and the Henriksson and Merton (1981) market-timing measure: *a*, results for data set 1; *b*, results for data set 2; *c*, results for France from data set 3.

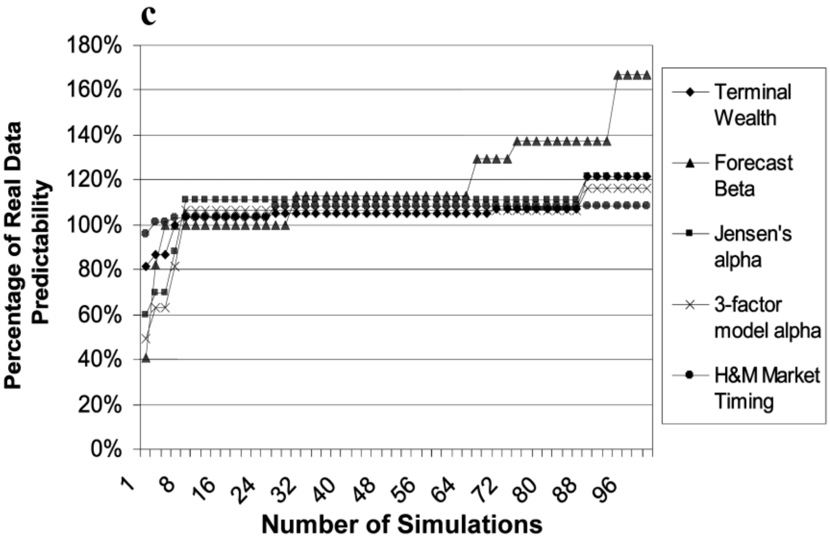

FIG. 1.—*Continued*

model to form a portfolio in step-ahead periods.[14] We assume that one has no particularly strong priors concerning the econometrician choice variables, except for priors implicitly imposed by each data set. For example, in data sets 1 and 2, one has the prior to consider a specific group of variables along with only a single U.S. asset. The prior on a fixed asset is relaxed in data set 3, when we include multiple assets.

We construct an out-of-sample time series of returns for each data set using the following recursive approach:

1. For all possible variable combinations, I ($I = 2^K - 1$ models [each model includes an intercept], where K is the number of predictive variables in each data set), all possible in-sample window lengths, W ($W = 10$, 15, and 20 years and an expanding window for data set 1 and $W = 5$ and 10 years and an expanding window for data sets 2 and 3), and all possible assets, A ($A = 1$ for data sets 1 and 2 and $A = 13$ for data set 3), we estimate, using OLS, a linear model of the form $r_\tau = \beta_I' X_{\tau-1,I} + \varepsilon_{\tau,I}$, where $X_{\tau-1,I}$ is an $(n_I + 1) \times 1$ vector of predictive variables, including a vector of ones for the intercept term, and r_τ is the excess return for asset A during in-sample period τ. We estimate the model in the in-sample period W and use the loadings on the

[14] Another approach to analyze the general robustness of time-series predictability would be to estimate full-period in-sample regressions and analyze subperiod stability of the predictive variables' betas. However, this approach would not endogenize the econometrician choice variables; i.e., it would not allow for real-time competition of predictive variables, assets, and estimation lengths.

predictive variables to form expected return estimates during the in-sample period. For each forecast series, we obtain a series of realized returns from the following trading strategy: go long asset *A* if the expected excess return estimate for that period is greater than zero, else invest in a T-bill. We then choose the best *W*, *A*, and *I* combination from the $W \times A \times (2^K - 1)$ total combinations on the basis of the average in-sample terminal wealth, standardized by the number of periods in *W*.[15]

2. Using the optimal model from above, we form a step-ahead out-of-sample forecast using the in-sample intercept and predictive variable loadings.
3. We then roll forward the in-sample end date by one period, repeat steps 1 and 2, and obtain a forecast for the next out-of-sample period. We repeat this process until the end of the out-of-sample period. Thus, for each data set, we obtain a single out-of-sample forecast series.
4. For the out-of-sample forecast series, we obtain a series of realized returns for the "active" portfolio from the following trading strategy: go long in the optimal asset *A* if the expected excess return estimate for that period is greater than zero, else invest in a T-bill. We do not allow leverage in the traded asset.

The results are presented in table 4. Across the three data sets, we find statistically significant predictability only in data set 2. The results for data set 2, in panel B, show that the active portfolio (assuming zero transaction costs) outperforms a buy-and-hold position in the S&P 500 by 15 basis points per month. Out of 348 months in the out-of-sample period, the active strategy trades 193 months. Although the active portfolio's raw mean is not that much greater than that of the S&P 500, the standard deviation is lower, resulting in more than double the Sharpe ratio. In addition, the active portfolio exhibits a significant forecast beta, market-timing statistic, and Jensen and Fama-French alphas.

We also endogenize various fixed transaction costs for the recursive experiment. We do this by altering the in-sample trading rule to "go long in asset *A* if the expected excess return is greater than zero plus the one-way transaction cost." Thus, under this setting, the optimal in-sample combination of assets, predictive variables, and window lengths is determined accounting for transaction costs and then applied to the step-ahead out-of-sample period. To form the active out-of-sample portfolio, we also require the expected return estimate to be greater than zero plus the transaction costs. We consider one-way transaction costs of 10, 30, and 50 basis points. The economic evidence of predictability for data set 2 is weakened after we account for the lowest

[15] We also examine the mean and Sharpe ratio as the in-sample objective function. The results (not reported, but available from the authors) are similar to those in the terminal wealth objective function.

level of one-way transaction costs. At the 10–basis point transaction cost level, the active portfolio now has a Sharpe ratio approximately the same as the buy-and-hold benchmark and has statistically insignificant values of the forecast beta and alphas, but retains a significant market-timing measure. For data sets 1 and 3, endogenizing transaction costs does not help in improving the performance of the active portfolio.

It is interesting to contrast our tests in this section with White's (2000) reality check approach. White develops a bootstrap approach to estimate the biases in statistical inference induced by data snooping.[16] The reality check generates a *p*-value for the null hypothesis that the performance of the best trading rule from a universe of trading rules is no better than a benchmark. Thus the reality check compares the performance of a fixed, best ex post model to the universe of all other models. In contrast, our recursive endogenization of predictive variables and estimation windows results in an optimal model that may vary each month over the out-of-sample period. Therefore, the reality check *p*-value does not directly apply. There are, however, ways in which White's reality check could be applied to future extensions of this paper. For example, as a method to incorporate learning, the best in-sample rule could be verified using White's *p*-value in a holdout confirmation period. If the rule's performance obtains a significant reality check *p*-value, only then would it be used in the out-of-sample period.

How much endogenizing is enough?—In light of the large number of potential parameters that researchers typically exogenously specify in order to conduct time-series predictability tests—we list 12 and test just three from our reality spectrum (table 1)—we ask the following question in this section: how much endogenizing is enough? If one finds predictability after endogenizing, say, one or two aspects, is that enough? If the goal were to be truly "real time," as in modeling the full spectrum of uncertainty that an actual investor faces, then *all* aspects would need to be endogenized. Obviously, this is not practical and probably impossible to implement. However, a recent positive trend (in our opinion) in the time-series literature has been the attempt to reduce ex post biases by endogenizing one or two aspects of real-time uncertainty. For example, papers have endogenized predictive variable selection (Pesaran and Timmermann 1995; Bossaerts and Hillion 1999; Pastor 2000;

[16] White's reality check is extremely flexible and robust; it allows for the use of numerous criteria to pick the best specification, allows for nonnormality and cross-correlation of returns, allows for time-series correlation of residuals, can be applied to single-period or recursive experiments, and exhibits desirable power characteristics, especially as the number of specifications evaluated grows large. White uses his reality check to examine the robustness of daily trading rules on the S&P 500. Sullivan et al. (2001) examine the robustness of calendar effects, and Sullivan et al. (1999) examine technical trading rules. Kosowski et al. (2003) apply the bootstrap to mutual fund manager performance. The results of the first three studies suggest that anomalous returns associated with calendar rules and technical trading rules are overstated. In the last study, using the bootstrap to control for the effects of luck in the distribution of fund alphas, the authors find that fund managers do exhibit performance persistence, and this persistence is particularly strong among growth fund managers.

TABLE 4 Endogenizing Predictive Variables and In-Sample Window Lengths via a Recursive Strategy

A. Quarterly Out-of-Sample Performance Results for Data Set 1, 1973(9)–1998(3)

	Mean Return (%)	Standard Deviation	Terminal Wealth ($)	Sharpe Ratio	β_f	Jensen's α (%)	FF α (%)	HM p-Value	HM $p_1 + p_2$	Active Trades	N
Buy-and-hold	3.36	8.08	18.99	.20		−.27	−.25*				99
Active: transaction cost =											
0%	3.05	5.46	17.08	.24	.24	.44	.33	.16	.97	56	99
.1%	3.03	5.62	16.53	.23	.18	.37	.31	.16	.97	56	99
.3%	2.49	5.47	9.95	.14	−.06	−.27	−.03	.03	.80	57	99
.5%	3.06	5.84	16.94	.23	.23	.31	.19	.17	.98	54	99

B. Monthly Out-of-Sample Performance Results for Data Set 2, 1964(1)–1992(12)

	Mean Return (%)	Standard Deviation	Terminal Wealth ($)	Sharpe Ratio	β_f	Jensen's α (%)	FF α (%)	HM p-Value	HM $p_1 + p_2$	Active Trades	N
Buy-and-hold	.92	4.36	17.55	.09		.03	.06				348
Active: transaction cost =											
0%	1.07	2.87	35.22	.19	.21**	.40***	.40***	.004	1.13	193	348
.1%	.82	3.21	14.44	.10	.16	.11	.08	.016	1.10	195	348
.3%	.73	3.00	10.92	.07	.21*	.05	.004	.051	1.05	178	348
.5%	.77	3.14	12.03	.08	.19**	.06	−.008	.046	1.06	170	348

C. Monthly Out-of-Sample Performance Results for Data Set 3, 1980(1)–1995(5)

	Mean Return (%)	Standard Deviation	Terminal Wealth ($)	Sharpe Ratio	Jensen's α (%)	FF α (%)	Average Number of Assets	Active Trades	*N*
Buy-and-hold	1.28	4.58	8.70	.15	.20	.08			185
Active: transaction cost =									
0%	1.31	4.93	8.92	.15	.23	.11	4.81	182	185
.1%	1.36	5.06	9.54	.15	.27	.11	4.64	182	185
.3%	1.18	4.99	6.91	.14	.22	−.04	4.49	183	185
.5%	1.07	5.14	5.63	.09	−.01	−.08	4.09	181	185

NOTE.—This table reports the results of strategies that endogenize predictive variables and in-sample window lengths via a recursive strategy. For each data set, we estimate, using OLS, a linear model of expected returns within an in-sample period for each predictive variable combination and window length. We use the loadings on the predictive variables and the estimated intercept to form expected return estimates during the in-sample period. We obtain a series of realized returns from the following trading strategy: go long asset *A* if the expected excess return estimate for that period is great than zero, else invest in a T-bill. We then choose the best combination from the $W \times A \times (2^K - 1)$ total models (where *W* is the in-sample window length, *A* is assets, and *K* is the number of predictive variables) based on the average in-sample terminal wealth, standardized by the number of periods in *W*. We examine 508 combination for data set 1 (1 asset × 4 windows × $[2^7-1]$ models), 1,533 combinations for data set 2 (1 asset ×3 windows × $[2^9-1]$ models), and 39,897 combinations (13 assets × 7 windows × $[2^{10}-1]$ models) for data set 3. Using the optimal in-sample combination, we form a step-ahead out-of-sample forecast using the in-sample intercept and predictive variable loadings. We then roll forward the in-sample end date by one period, find again the best in-sample combination, and obtain a forecast for the next out-of-sample period. We repeat this process until the end of the out-of-sample period. For the out-of-sample forecast series, we obtain a series of realized returns for the "active" portfolio from the following trading strategy: go long in the optimal asset *A* if the expected excess return estimate for that period is great than zero, else invest in a T-bill. We report the out-of-sample mean and standard deviation of returns for the active portfolio and the buy-and-hold benchmark strategy. For data sets 1 and 2, the buy-and-hold is a constant position in the S&P 500. For data set 3, the buy-and-hold is an equally weighted portfolio of the 13 country assets. Terminal wealth is the total wealth at the end of the out-of-sample period of investing one dollar at the beginning. We also report a Sharpe ratio, Jensen's alpha, and Fama-French (1993) three-factor alpha. For data sets 1 and 2, we also report the Henriksson and Merton (1981) market-timing statistics and the forecast beta, which provides a measure of overall out-of-sample fit and is calculated by regressing the monthly realized return on the forecasted return $r_\tau = \alpha + \beta_f r_{\text{forecast},\tau} + \varepsilon_\tau$. "Active trades" reports the number of periods that the active portfolio invests in the risky asset. *N* shows the number of out-of-sample periods.

* Significant at the 10% level.

** Significant at the 5% level.

*** Significant at the 1% level.

Ait-Sahalia and Brandt 2001; Avramov 2002; Cremers 2002), in-sample window length (Pesaran and Timmermann 2002), and statistical model selection (Pesaran and Timmermann 1995).[17]

In table 5 we report the percentage of out-of-sample forecasts for which the null is rejected in experiments in which we endogenize one or more parameters, while exogenously looping over other parameter values. In panel A we endogenize variable selection via six statistical model selection criteria and loop over exogenous values of in-sample windows lengths and assets,[18] in panel B we endogenize window length and loop over exogenous values of model selection criteria and assets, in panel C we endogenize model selection criteria and loop over exogenous values of window lengths and assets, and in panel D, we endogenize both model selection criteria and windows and examine variations across assets (for data set 3). We use the same recursive methodology as in Section III.*B*, but now we first endogenize one aspect via a terminal wealth objective function and then generate other out-of-sample portfolios by varying the values of the exogenous choice variables.[19] We do not use transaction costs in this subsection.

The results in panels A, B, and C suggest that exogenous parameter selection has a large effect on how often one rejects the null *even when other aspects are endogenized*. For example, in panel A, after we endogenize variable selection, data set 1 experiences rejections of the null in 4.2%–20.8% of the exogenous model selection and window length specifications, depending on which performance measure is used. We observe similar patterns for all three data sets in panels A, B, and C, with data set 2 experiencing the most rejections of the null across panels and performance measures. Across the data sets and panels, there does not appear to be any consistent pattern in which type of model selection or window length rejects the null. Finally, in panel D, after endogenizing both windows and model selection criteria for the 13 country assets, we find that only the French index survives this process, generating significant alphas and market timing but not a significant forecast beta.

Pesaran and Timmermann (1995) present results (in their table 3) in which they endogenize variable selection using eight model selection criteria and an expanding window. Across the eight model selection criteria, they find evidence of predictability using a joint test on capital asset pricing model alphas

[17] Arbitrage pricing theory papers such as Roll and Ross (1980) and Dhrymes, Friend, and Gultekin (1984) use factor analysis to extract priced factors from historical returns. Thus these papers can also be viewed as endogenizing predictive variables.

[18] The six criteria are Akaike's information criterion, Schwarz's Bayesian information criterion, Sawa's Bayesian information criterion, Amemiya's prediction criteria, adjusted R^2, and a model that uses all predictive variables.

[19] For example, consider panel B of table 5, in which we endogenize window length for each selection criterion and asset. During the in-sample period, for each selection criterion, we find the variable combination with the highest value of the selection criteria for each window length. Using the best model for each window, we then find the window that results in the highest average terminal wealth. Thus we arrive at the best window length for each of the six selection criteria and asset(s).

but conclude that most of their evidence is concentrated in the 1970s. Our results suggest that their findings may be dependent on the choice of window length: in panel A of table 5, for data set 2, we find that between 38.9% and 72.2% of models reject the null as we vary in-sample window length (five years, 10 years, and expanding). In panel B, where we let the data pick the best window, we find that 100% of the models reject the null for three test statistics (Jensen's alpha, the Fama-French alpha, and the market-timing measure), but no models reject the null using the forecast beta. Pesaran and Timmermann also examine a "hyper-selectivity" criterion in which they endogenize model selection criteria (p. 1226). Panel C of table 5 provides the reader with a sense of how sensitive their hyper-selection procedure is to window length variations. Across three window lengths, we find that between one and two specifications result in predictability across the various test statistics.

Overall, this section shows that in experiments in which one aspect is endogenized, the rejection of the null is heavily dependent on the value of other exogenously specified choice variables. And again, since these choice variables vary across the range of possible values, it appears unlikely that one would posses an ex ante prior on the successful models.

IV. Conclusion

Previous papers on data snooping have shown the dangers of using an *in-sample* methodology to test predictive models. One of the solutions suggested by these papers to control for in-sample test size problems is to use holdout periods (out-of-sample tests) to validate predictive models. The main contribution of this paper is to show that these out-of-sample tests used in the time-series literature suffer from test size problems related to exogenous parameter specification occurring when the researcher has knowledge of the full data set.

Our results carry implications for the growing numbers of conditional asset pricing studies that exogenously choose lagged factors on the basis of their ability to forecast the general market and employ loadings from these variables in cross-sectional tests (Ferson and Harvey 1999; Lettau and Ludvigson 2001; Santos and Veronesi 2006). Similarly, the results have implications for the rapidly growing Bayesian predictability literature that typically chooses an exogenous set of predictive variables and shows how "parameter uncertainty" can lead to important changes in investors' allocations to stocks (Kandel and Stambaugh 1996; Barberis 2000).[20] In both the Kandel and Stambaugh and

[20] In related work, Lewellen and Shanken (2001) argue that the Bayesian learning of economic agents can generate ex post predictable patterns that are ex ante rational and therefore not real-time tradable opportunities. In this case, predictability is just an ex post illusion. For example, suppose that you know that the time series of stock returns is mean-reverting. In real time, you still do not know if stock prices will be higher or lower next period because you do not know

TABLE 5 **Rejection Rates of the Null for Combinations of Exogenously and Endogenously Specified Parameters**

				Percentage of Out-of-Sample Forecasts Rejecting the Null			
Data Set	Endogenized Parameter	Exogenous Specifications	Total Number of Specifications	Forecast Beta $\beta_f > 0$ ($p_{\beta f} \leq .05$)	Jensen's α $\alpha_j > \alpha_{j,bh}$ ($p_{\alpha} \leq .05$)	FF α $\alpha_{\mathrm{ff}} > \alpha_{\mathrm{ff},bh}$ ($p_{\alpha ff} \leq .05$)	Market Timing $\mathrm{HM}_{p1+p2} > 1$ ($\mathrm{HM}_p \leq .05$)
		A. Endogenized Variable Selection					
1	Variable selection	6 selection criteria × 4 windows	24	20.8	12.5	4.2	4.2
2	Variable selection	6 selection criteria × 3 windows	18	66.7	44.4	38.9	72.2
3	Variable selection	6 selection criteria × 3 windows × 13 assets	234	1.3	1.7	.4	5.1
		B. Endogenized In-Sample Window Lengths					
1	Window	6 selection criteria	6	0	33.3	0	0
2	Window	6 selection criteria	6	0	100	100	100
3	Window	6 selection criteria × 13 assets	78	0	1.3	0	2.6
		C. Endogenized Statistical Model Selection Criteria					
1	Selection criteria	4 windows	4	0	0	0	0
2	Selection criteria	3 windows	3	66.7	33.3	33.3	66.7
3	Selection criteria	3 windows × 13 assets	39	0	2.6	0	5.1

	D. Endogenized Window and Statistical Model Selection Criteria						
3	Window and selection criteria	13 assets	13	0	7.7	7.7	7.7

NOTE.—This table presents the number of cases for which the null hypothesis of no predictability is rejected using various performance measures for monthly/quarterly out-of-sample forecasts based on combinations of exogenous and endogenous parameter specification. The coefficient estimate of the slope (β_f) provides a measure of overall out-of-sample fit and is calculated by regressing the monthly realized return on the forecasted return $r_\tau = \alpha + \beta_f r_{\text{forecast},\tau} + \varepsilon_\tau$.

the Barberis papers, predictability emanates from the dividend yield, making their results conditional on that specific model.[21] Our evidence suggests that caution needs to be exercised in interpreting conclusions from such studies given their assumptions on a specific model of predictability.

In summary, the simulations and real-time methodology used in this paper do not suggest an alternative model of the factors that drive aggregate market returns. The power to detect real-time market predictability may be increased by incorporating other aspects of uncertainty that we have not considered, such as other predictive variables, different assets, multiple return horizons, nonlinear models, different forms of learning, and other changes. However, our results suggest that it is critically important to control for the degrees of freedom available to a researcher in specifying these parameters in out-of-sample tests. To that end, this paper provides two methods to control for biases arising from such parameter freedom. First, simulations, such as those found in this paper, can be used to adjust the null hypothesis of no predictability in out-of-sample tests. Second, the researcher could use variants of this paper's recursive tests to endogenize parameter selection. Thus these adjustments may allow for a better identification of the true factors that drive aggregate market returns.

References

Ait-Sahalia, Y., and M. Brandt. 2001. Variable selection for portfolio choice. *Journal of Finance* 56:1297–1351.

Allen, F., and R. Karjalainen. 1999. Using genetic algorithms to find technical trading rules. *Journal of Financial Economics* 51:245–71.

Ang, A., and G. Bekaert. 2006. Stock return predictability: Is it there? Working paper, Columbia University.

Avramov, D. 2002. Stock return predictability and model uncertainty. *Journal of Financial Economics* 64:423–58.

Barber, B., R. Lehavy, M. McNichols, and B. Trueman. 2001. Can investors profit from the prophets?: Security analysts' recommendations and stock returns. *Journal of Finance* 56: 531–63.

Barber, B., and T. Odean. 2000. Trading is hazardous to your wealth: The common stock investment performance of individual investors. *Journal of Finance* 55:773–806.

Barberis, N. 2000. Investing for the long run when returns are predictable. *Journal of Finance* 55:225–64.

Black, F. 1993a. Beta and return. *Journal of Portfolio Management* 20:8–18.

———. 1993b. Estimating expected return. *Financial Analysts Journal* 49:36–38.

Bossaerts, P., and P. Hillion. 1999. Implementing statistical criteria to select return forecasting models: What do we learn? *Review of Financial Studies* 12:405–28.

Breen, W., L. Glosten, and R. Jagannathan. 1989. Economic significance of predictable variations in stock index returns. *Journal of Finance* 44:1177–89.

the true mean of the distribution. Nonetheless, a pattern of mean reversion is easily detected ex post relative to the sample mean.

[21] In related non-Bayesian work, Goyal and Welch (2003) document substantial in-sample predictability in the time series of stock index returns based on dividend yields, but they find no evidence of out-of-sample forecastability. They attribute the difference in performance between in- and out-of-sample predictability to parameter instability, i.e., a time-varying correlation between expected returns and dividend yield.

Brown, S., W. Goetzmann, and A. Kumar. 1998. The Dow theory: William Peter Hamilton's track record reconsidered. *Journal of Finance* 53:1311–33.

Brown, S., W. Goetzmann, and S. Ross. 1995. Survival. *Journal of Finance* 50:853–73.

Campbell, J. 1987. Stock returns and the term structure. *Journal of Financial Economics* 18: 373–99.

Campbell, J., and R. Shiller. 1988a. The dividend-price ratio and expectations of future dividends and discount factors. *Review of Financial Studies* 1:195–228.

———. 1988b. Stock prices, earnings, and expected dividends. *Journal of Finance* 43:661–76.

Carhart, M. 1997. On persistence in mutual fund performance. *Journal of Finance* 52:57–82.

Christopherson, J., W. Ferson, and D. Glassman. 1998. Conditioning manager alphas on economic information: Another look at the persistence of performance. *Review of Financial Studies* 11: 111–42.

Cochrane, J. 1991. Production based asset pricing and the link between stock returns and macroeconomic fluctuations. *Journal of Finance* 46:209–38.

———. 1999. Portfolio advice for a multifactor world. *Federal Reserve Bank of Chicago Economic Perspectives* 23, no. 3:59–78.

Conrad, J., M. Cooper, and G. Kaul. 2002. Value versus glamour. *Journal of Finance* 58:1969–95.

Cooper, M., R. Gutierrez, and W. Marcum. 2005. On the predictability of stock returns in real time. *Journal of Business* 78 (March): 469–99.

Coval, J., and T. Shumway. 2001. Is sound just noise? *Journal of Finance* 56, no. 5:1887–1910.

Cremers, M. 2002. Stock return predictability: A Bayesian model selection perspective. *Review of Financial Studies* 15:1223–49.

Denton, F. 1985. Data mining as an industry. *Review of Economics and Statistics* 67:124–27.

Desai, H., and P. Jain. 1995. An analysis of the recommendations of the "superstar" money managers at Barron's annual roundtable. *Journal of Finance* 50:1257–74.

Dhrymes, P., I. Friend, and N. Gultekin. 1984. A critical reexamination of the empirical evidence on the arbitrage pricing theory. *Journal of Finance* 39:323–46.

Fama, E. 1991. Efficient capital markets II. *Journal of Finance* 46:1575–1643.

Fama, E., and K. French. 1988. Dividend yields and expected stock returns. *Journal of Financial Economics* 22:3–25.

———. 1989. Business conditions and expected returns on stocks and bonds. *Journal of Financial Economics* 25:23–49.

———. 1993. Common risk factors in the returns of stocks and bonds. *Journal of Financial Economics* 33:3–56.

Ferson, W., and C. Harvey. 1991. The variation of economic risk premiums. *Journal of Political Economy* 99:385–415.

———. 1993. The risk and predictability of international equity returns. *Review of Financial Studies* 6:527–66.

———. 1999. Conditioning variables and the cross section of stock returns. *Journal of Finance* 54:1325–60.

Ferson, W., S. Sarkissian, and T. Simin. 2003. Spurious regressions in financial economics? *Journal of Finance* 58:1393–1413.

Foster, D., T. Smith, and R. Whaley. 1997. Assessing goodness-of-fit of asset pricing models: The distribution of the maximal R^2. *Journal of Finance* 52:591–607.

Goetzmann, W., and P. Jorion. 1999. Global stock markets in the twentieth century. *Journal of Finance* 54:953–80.

Goyal, A., and I. Welch. 2003. Predicting the equity premium with dividend ratios. *Management Science* 49, no. 5:639–54.

Hau, H. 2001. Location matters: An examination of trading profits. *Journal of Finance* 56: 1959–83.

Henriksson, R., and R. Merton. 1981. On market timing and investment performance. II. Statistical procedure for evaluating forecasting skills. *Journal of Business* 54:513–33.

Hirshleifer, D., and T. Shumway. 2001. Good day sunshine: Stock returns and the weather. *Journal of Finance* 58:1009–32.

Hodrick, R. 1992. Dividend yields and expected stock returns: Alternative procedures for inference and measurement. *Review of Financial Studies* 5:357–86.

Kandel, S., and R. Stambaugh. 1996. On the predictability of stock returns: An asset allocation perspective. *Journal of Finance* 51:385–424.

Keim, D., and R. Stambaugh. 1986. Predicting returns in the stock and bond markets. *Journal of Financial Economics* 17:357–90.

Kosowski, R., A. Timmermann, H. White, and R. Wermers. 2003. Can mutual fund stars really pick stocks? New evidence from a bootstrap analysis. Working paper, University of California, San Diego.

Lamont, O. 1998. Earnings and expected returns. *Journal of Finance* 53:1563–87.

Lettau, M., and S. Ludvigson. 2001. Consumption, aggregate wealth and expected stock returns. *Journal of Finance* 56:815–49.

Lewellen, J. 1999. The time-series relations among expected return, risk, and book-to-market. *Journal of Financial Economics* 54:5–43.

Lewellen, J., and J. Shanken. 2001. Learning, asset-pricing tests, and market efficiency. *Journal of Finance* 57:1113–45.

Lo, A., and A. MacKinlay. 1990. Data-snooping biases in tests of financial asset pricing models. *Review of Financial Studies* 3:431–67.

———. 1997. Maximizing predictability in the stock and bond markets. *Macroeconomic Dynamics* 1:102–34.

Metrick, A. 1999. Performance evaluation with transactions data: The stock selection of investment newsletters. *Journal of Finance* 54:1743–75.

Pastor, L. 2000. Portfolio selection and asset pricing models. *Journal of Finance* 55:179–223.

Pesaran, M., and A. Timmermann. 1995. Predictability of stock returns: Robustness and economic significance. *Journal of Finance* 50:1201–28.

———. 2002. Model instability and choice of observation window. Working paper, University of California, San Diego.

Pirinsky, C. 2001. Are financial institutions better investors? Manuscript, Ohio State University.

Pontiff, J., and L. Schall. 1998. Book-to-market ratios as predictors of market returns. *Journal of Financial Economics* 49:141–60.

Roll, R., and S. Ross. 1980. An empirical investigation of the arbitrage pricing theory. *Journal of Finance* 35:1073–1103.

Santos, T., and P. Veronesi. 2006. Labor income and predictable stock returns. *Review of Financial Studies* 19:1–44.

Shiller, R. 1984. Stock prices and social dynamics. *Brookings Papers on Economic Activity*, no. 2, pp. 457–98.

Sullivan, R., A. Timmermann, and H. White. 1999. Data-snooping, technical trading rule performance, and the bootstrap. *Journal of Finance* 54:1647–91.

———. 2001. Dangers of data-driven inferences: The case of calendar effects. *Journal of Econometrics* 105, no. 1:249–86.

Swanson, N., and H. White. 1997. A model selection approach to real-time macroeconomic forecasting using linear models and artificial neural networks. *Review of Economics and Statistics* 79:540–50.

Wermers, R. 2000. Mutual fund performance: An empirical decomposition into stock-picking talent, style, transactions costs, and expenses. *Journal of Finance* 55:1655–95.

White, H. 2000. A reality check for data snooping. *Econometrica* 68, no. 5:1097–1126.