… and the Cross-Section of Expected Returns

Campbell R. Harvey
Duke University, National Bureau of Economic Research

Yan Liu
Texas A&M University

Heqing Zhu
The University of Oklahoma

Hundreds of papers and factors attempt to explain the cross-section of expected returns. Given this extensive data mining, it does not make sense to use the usual criteria for establishing significance. Which hurdle should be used for current research? Our paper introduces a new multiple testing framework and provides historical cutoffs from the first empirical tests in 1967 to today. A new factor needs to clear a much higher hurdle, with a t-statistic greater than 3.0. We argue that most claimed research findings in financial economics are likely false. (*JEL* C12, C52, G12)

Received October 22, 2014; accepted June 15, 2015 by Editor Andrew Karolyi.

Over forty years ago, one of the first tests of the capital asset pricing model (CAPM) found that the market beta was a significant explanator of the cross-section of expected returns. The reported t-statistic of 2.57 in Fama and MacBeth (1973, Table III) comfortably exceeded the usual cutoff of 2.0. However, since that time, hundreds of papers have tried to explain the cross-section of expected returns. Given the known number of factors that have been tried and the reasonable assumption that many more factors have been tried but did not make it to publication, the usual cutoff levels for statistical significance

We would like to thank the Editor (Andrew Karolyi) and three anonymous referees for their detailed and thoughtful comments. We would also like to thank Viral Acharya, Jawad Addoum, Tobias Adrian, Andrew Ang, Ravi Bansal, Mehmet Beceren, Itzhak Ben-David, Bernard Black, Jules van Binsbergen, Oliver Boguth, Tim Bollerslev, Alon Brav, Ian Dew-Becker, Robert Dittmar, Jennifer Conrad, Michael Cooper, Andres Donangelo, Gene Fama, Wayne Ferson, Ken French, Simon Gervais, Bing Han, John Hand, Abby Yeon Kyeong Kim, Lars-Alexander Kuehn, Sophia Li, Harry Markowitz, Kyle Matoba, David McLean, Marcelo Ochoa, Peter Park, Lubos Pastor, Andrew Patton, Lasse Pedersen, Tapio Pekkala, Jeff Pontiff, Ryan Pratt, Tarun Ramadorai, Alexandru Rosoiu, Tim Simin, Avanidhar Subrahmanyam, Ivo Welch, Basil Williams, Yuhang Xing, Josef Zechner, and Xiaofei Zhao, as well as seminar participants at the 2014 New Frontiers in Finance Conference at Vanderbilt University, the 2014 Inquire Europe-UK meeting in Vienna, the 2014 WFA meetings, and seminars at Duke University, Texas A&M University, Baylor University, University of Utah, and Penn State University. Our data are available for download and resorting. The main table includes full citations and Web addresses to each of the cited articles. See http://faculty.fuqua.duke.edu/~charvey/Factor-List.xlsx. Supplementary data can be found on *The Review of Financial Studies* web site. Send correspondence to Campbell R. Harvey, Fuqua School of Business, Duke University, Durham, NC 27708; telephone: 919-660-7768. E-mail: cam.harvey@duke.edu.

doi:10.1093/rfs/hhv059 Advance Access publication October 9, 2015

may not be appropriate. We present a new framework that allows for multiple tests and derive recommended statistical significance levels for current research in asset pricing.

We begin with 313 papers published in a selection of journals that study cross-sectional return patterns. We provide recommended test thresholds from the first empirical tests in 1967 to present day. We also project minimum t-statistics through 2032, assuming the rate of "factor production" remains the same as the last ten years. We present a taxonomy of historical factors, as well as definitions.[1]

Our research is related to a recent paper by McLean and Pontiff (2015), who argue that certain stock market anomalies are less anomalous after being published.[2] Their paper tests the statistical biases emphasized in Leamer (1978), Ross (1989), Lo and Mackinlay (1990), Fama (1991), and Schwert (2003).

Our paper also adds to the recent literature on biases and inefficiencies in cross-sectional regression studies. Lewellen, Nagel, and Shanken (2010) critique the usual practice of using cross-sectional R^2s and pricing errors to judge success and show that the explanatory power of many previously documented factors are spurious. Our work focuses on evaluating the statistical significance of a factor given the previous tests on other factors. Our goal is to use a multiple testing framework to both re-evaluate past research and to provide a new benchmark for current and future research.

We tackle multiple hypothesis testing from the frequentist perspective. Bayesian approaches to multiple testing and variable selection also exist.[3] However, the high dimensionality of the problem combined with the fact that we do not observe all the factors that have been tried poses a big challenge for Bayesian methods. While we propose a frequentist approach to overcome this missing data issue, it is unclear how to do this in the Bayesian framework. Nonetheless, we provide a detailed discussion of Bayesian methods in paper.

Multiple testing has only recently gained traction in the finance literature. For the literature on multiple testing corrections for data snooping biases, see Sullivan, Timmermann, and White (1999, 2001) and White (2000). For research on data snooping and variable selection in predictive regressions, see Foster, Smith, and Whaley (1997), Cooper and Gulen (2006), and Lynch and Vital-Ahuja (2012). For applications of multiple testing approach in the finance literature, see, for example, Shanken (1990), Ferson and Harvey (1999),

[1] We also provide a link to a file with full references and Web addresses to the original articles: http://faculty.fuqua.duke.edu/~charvey/Factor-List.xlsx.

[2] Other recent papers that systematically study the cross-sectional return patterns include those by Subrahmanyam (2010) and Green, Hand, and Zhang (2013a, 2013b). Other papers that study anomaly discoveries and investor actions include those by Edelen, Ince, and Kadlec (2014) and Liu et al. (2014).

[3] See Jefferys and Berger (1992), Scott and Berger (2006), and Scott (2009).

Boudoukh et al. (2007), and Patton and Timmermann (2010). More recently, a multiple testing connection has been used to study technical trading and mutual fund performance, see, for example, Barras, Scaillet, and Wermers (2010), Bajgrowicz and Scaillet (2012), and Kosowski et al. (2006). Conrad, Cooper, and Kaul (2003) point out that data snooping accounts for a large proportion of the return differential between equity portfolios that are sorted by firm characteristics. Bajgrowicz, Scaillet, and Treccani (2013) show that multiple testing methods help eliminate a large proportion of spurious jumps detected using conventional test statistics for high-frequency data. Holland, Basu, and Sun (2010) emphasize the importance of multiple testing in accounting research. Our paper is consistent with the theme of this literature.

There are limitations to our framework. First, should all factor discoveries be treated equally? We think no. A factor derived from a theory should have a lower hurdle than a factor discovered from a purely empirical exercise. Economic theories are based on a few economic principles and, as a result, there is less room for data mining. Nevertheless, whether suggested by theory or empirical work, a t-statistic of 2.0 is too low. Second, our tests focus on unconditional tests. While the unconditional test might consider the factor marginal, it is possible that this factor is very important in certain economic environments and not important in other environments. These two caveats need to be taken into account when using our recommended significance levels for current asset pricing research.

While our focus is on the cross-section of equity returns, our message applies to many different areas of finance. For instance, Frank and Goyal (2009) investigate around thirty variables that have been documented to explain the capital structure decisions of public firms. Welch and Goyal (2008) examine the performance of a dozen variables that have been shown to predict market excess returns. Novy-Marx (2014) proposes unconventional variables to predict anomaly returns. These three applications are ideal settings to employ multiple testing methods.

1. The Search Process

Our goal is not to catalog every asset pricing paper ever published. We narrow the focus to papers that propose and test new factors. For example, Sharpe (1964), Lintner (1965), and Mossin (1966) all theoretically proposed (at roughly the same time), a single-factor model—the capital asset pricing model (CAPM). Following Fama and MacBeth (1973), there are hundreds of papers that test the CAPM. We include the theoretical papers, as well as the first paper to provide test statistics. We do not include the hundreds of papers that test the CAPM in different contexts, for example, various international markets and different time periods. We do, however, include papers, such as Kraus and Litzenberger (1976), who test the market factor, as well as one additional risk factor linked to the market factor.

Sometimes different papers propose different empirical proxies for the same type of economic risk. Although they may look similar from a theoretical standpoint, we still include them. An example is the empirical proxies for idiosyncratic financial constraints risk. While Lamont, Polk, and Saa-Requejo (2001) use the Kaplan and Zingales (1997) index to proxy for firm-level financial constraints, Whited and Wu (2006) estimate their own constraint index based on the first-order conditions of firms' optimization problem. We include both even though they are likely highly correlated.

Since our focus is on factors that can broadly explain return patterns, we omit papers that focus on a small group of stocks or a short period of time. This will, for example, exclude a substantial amount of empirical corporate finance research that studies event-driven return movements.[4]

Certain theoretical models lack immediate empirical content. Although they could be empirically relevant once suitable proxies are constructed, we choose to exclude them.

With these rules in mind, we narrow our search to generally the top journals in finance, economics, and accounting. To include the most recent research, we search for working papers on the Social Science Research Network (SSRN). Working papers pose a challenge because there are thousands of them, and they have not been subjected to peer review. We choose a subset of papers that we suspect are in review at top journals, have been presented at top conferences, or are due to be presented at top conferences. We end with 63 working papers. In total, we focus on 313 articles, among which are 250 published articles. We catalogue 316 different factors.[5]

Our collection of 316 factors likely underrepresents the factor population. First, we generally only consider top journals. Second, we are selective in choosing only a handful of working papers. Third, sometimes there are many variants of the same characteristic, and we usually only include the most representative ones. Fourth, and perhaps most importantly, we should be measuring the number of factors tested (which is unobservable)—that is, we do not observe the factors that were tested but that failed to pass the usual significance levels and were never published (see Fama 1991). Our multiple testing framework tries to account for this possibility.

---

4 See Kothari and Warner (2007) for a survey on event studies. More specifically, three criteria help differentiate our risk factors from event signals in corporate finance. First, while we are generally considering returns realized at the monthly or lower frequency intervals for risk factors, it is routine for event studies to consider daily or even higher frequency returns. Second, portfolio sorts based on risk factors typically cover the entire cross-section of stocks, whereas event studies usually focus on a much smaller group of securities that are affected by the event signal. Finally, portfolio sorts based on risk factors are usually repeated at a fixed time interval, whereas events may happen sporadically.

5 As already mentioned, some of these factors are highly correlated. For example, we include four versions of idiosyncratic volatility, that is, Fama and MacBeth (1973), Ali, Hwang, and Trombley (2003), Ang et al. (2006), and Fu (2009).

2. Factor Taxonomy

To facilitate our analysis, we group the factors into different categories. We start with two broad categories: "common" and individual firm "characteristics." "Common" means the factor can be viewed as a proxy for a common source of risk. Risk exposure to this factor or its innovations is supposed to help explain cross-sectional return patterns. "Characteristics" means the factor is specific to the security or portfolio. A good example is Fama and MacBeth (1973). While the beta against the market return is systematic (exposure to a common risk factor), the standard deviation of the market model residual is not based on a common factor—it is a property of the individual firm, that is, it is an idiosyncratic characteristic.

Strictly speaking, a risk factor should be a variable that has unpredictable variation through time. Moreover, assets' risk exposures to this factor need to be able to explain the cross-sectional return patterns. Based on these criteria, individual firm characteristics should not qualify as risk factors because characteristics are preknown and have limited time-series variation. However, we interpret firm characteristics in a broader sense. If a certain firm characteristic is found to be correlated with the cross-section of expected returns, a long-short portfolio can usually be constructed to proxy for the underlying unknown risk factor. It is this unknown risk factor that we have in mind when we classify particular firm characteristics as risk factors.

Based on the unique properties of the proposed factors, we further divide the "common" and "characteristics" groups into finer categories. In particular, we divide "common" into "financial," "macro," "microstructure," "behavioral," "accounting," and "other." We divide "characteristics" into the same categories, except we omit the "macro" classification, which is common, by definition. The following table provides further details on the definitions of these subcategories and gives examples for each.

3. Adjusted *t*-statistics in Multiple Testing

3.1 Why multiple testing?

Given that so many papers have attempted to explain the same cross-section of expected returns, statistical inference should not be based on a "single" test perspective. Our goal is to provide guidance as to the appropriate significance level using a multiple testing framework. When just one hypothesis is tested, we use the term "individual test," "single test," and "independent test" interchangeably.[6]

Strictly speaking, different papers study different sample periods and hence focus on different cross-sections of expected returns. However, the bulk of the papers we consider have substantial overlapping sample periods. Also, if one

[6] The last term should not be confused with any sort of stochastic independence.

Table 1
Factor classification

Risk type		Description	Examples
Common (113)	**Financial** (46)	Proxy for aggregate financial market movement, including market portfolio returns, volatility, squared market returns, among others	Sharpe (1964): market returns; Kraus and Litzenberger (1976): squared market returns
	Macro (40)	Proxy for movement in macroeconomic fundamentals, including consumption, investment, inflation, among others	Breeden (1979): consumption growth; Cochrane (1991): investment returns
	Microstructure (11)	Proxy for aggregate movements in market microstructure or financial market frictions, including liquidity, transaction costs, among others	Pastor and Stambaugh (2003): market liquidity; Lo and Wang (2006): market trading volume
	Behavioral (3)	Proxy for aggregate movements in investor behavior, sentiment or behavior-driven systematic mispricing	Baker and Wurgler (2006): investor sentiment; Hirshleifer and Jiang (2010): market mispricing
	Accounting (8)	Proxy for aggregate movement in firm-level accounting variables, including payout yield, cash flow, among others	Fama and French (1992): size and book-to-market; Da and Warachka (2009): cash flow
	Other (5)	Proxy for aggregate movements that do not fall into the above categories, including momentum, investors' beliefs, among others	Carhart (1997): return momentum; Ozoguz (2009): investors' beliefs
Characteristics (202)	**Financial** (61)	Proxy for firm-level idiosyncratic financial risks, including volatility, extreme returns, among others	Ang et al. (2006): idiosyncratic volatility; Bali, Cakici, and Whitelaw (2011): extreme stock returns
	Microstructure (28)	Proxy for firm-level financial market frictions, including short sale restrictions, transaction costs, among others	Jarrow (1980): short sale restrictions; Mayshar (1981): transaction costs
	Behavioral (3)	Proxy for firm-level behavioral biases, including analyst dispersion, media coverage, among others	Diether, Malloy, and Scherbina (2002): analyst dispersion; Fang and Peress (2009): media coverage
	Accounting (87)	Proxy for firm-level accounting variables, including PE ratio, debt-to-equity ratio, among others	Basu (1977): PE ratio; Bhandari (1988): debt-to-equity ratio
	Other (24)	Proxy for firm-level variables that do not fall into the above categories, including political campaign contributions, ranking-related firm intangibles, among others	Cooper, Gulen, and Ovtchinnikov (2010): political campaign contributions; Edmans (2011): intangibles

The numbers in parentheses represent the number of factors identified. See Table 6 and http://faculty.fuqua.duke.edu/~charvey/Factor-List.xlsx.

believes that cross-sectional return patterns are stationary, then these papers are studying roughly the same cross-section of expected returns.

We want to emphasize that there are many forces that make our guidance lenient; that is, a credible case can be made for an even higher threshold for discovery. We have already mentioned that we only sample a subset of research papers and the "publication bias/hidden tests" issue (i.e., it is difficult to publish a nonresult).[7] However, there is another publication bias that is more subtle. In many scientific fields, replication studies routinely appear in top journals. That is, a factor is discovered, and others try to replicate it. In finance and economics, it is very difficult to publish replication studies. Hence, there is a bias towards publishing "new" factors rather than rigorously verifying the existence of discovered factors.

There are two ways to deal with the bias introduced by multiple testing: out-of-sample validation and using a statistical framework that allows for multiple testing.[8] When feasible, out-of-sample testing is the cleanest way to rule out spurious factors. In their study of anomalies, McLean and Pontiff (2015) take the out-of-sample approach. Their results show a degradation of performance of identified anomalies after publication, which is consistent with the statistical bias. It is possible that this degradation is larger than they document. In particular, they drop 12 of their 97 anomalies because they could not replicate the in-sample performance of published studies. Given that these nonreplicable anomalies were not even able to survive routine data revisions, they are likely to be insignificant strategies, either in-sample or out-of-sample. The degradation from the original published "alpha" is 100% for these strategies, which would lead to a higher average rate of degradation for their strategies.

While the out-of-sample approach has many strengths, it has one important drawback: it cannot be used in real time. To make real time assessments in the out-of-sample approach, it is common to hold out some data. However, this is not genuine out-of-sample testing as all the data are observable to researchers. A real out-of-sample test requires data in the future. In contrast to many tests in the physical sciences (where new data can be created for an experiment), we often need years of data to do an out-of-sample test. We pursue the multiple testing framework because it yields immediate guidance on whether a discovered factor is real.

3.2 A multiple testing framework

In statistics, multiple testing refers to simultaneous testing of more than one hypothesis. The statistics literature was aware of this multiplicity problem at

[7] See Rosenthal (1979) for one of the earliest and most influential works on publication bias.

[8] Another approach to test factor robustness is to look across multiple asset classes. This approach has been followed in several recent papers, for example, Asness et al. (2013) and Koijen et al. (2012).

Table 2
Contingency table in testing M hypotheses

Panel A: An example

	Unpublished	Published	Total
Truly insignificant	500	50	550
Truly significant	100	50	150
Total	600	100(R)	700(M)

Panel B: The testing framework

	H_0 not rejected	H_0 rejected	Total
H_0 true	$N_{0\|a}$	$N_{0\|r}$	M_0
H_0 false	$N_{1\|a}$	$N_{1\|r}$	M_1
Total	$M-R$	R	M

Panel A shows a hypothetical example for factor testing. Panel B presents the corresponding notation in a standard multiple testing framework.

least 60 years ago.[9] Early generations of multiple testing procedures focus on the control of the family-wise error rate (see Section 4.3.1). More recently, increasing interest in multiple testing from the medical literature has spurred the development of methods that control the false discovery rate (see Section 4.3.2). Multiple testing is an active research area in both the statistics and the medical literature.[10]

Despite the rapid development of multiple testing methods, they have not attracted much attention in the finance literature. Moreover, most of the research that does involve multiple testing focuses on the Bonferroni adjustment,[11] which is known to be too stringent. Our paper aims to fill this gap.

First, we introduce a hypothetical example to motivate a more general framework. In Table 2, we categorize the possible outcomes of a multiple testing exercise. Panel A displays an example of what the literature could have discovered, and panel B notationalizes panel A to ease our subsequent definition of the general type I error rate—the chance of making at least one false discovery or the expected fraction of false discoveries.

Our example in panel A assumes 100 published factors (denoted as R). Among these factors, suppose 50 are false discoveries and the rest are real ones. In addition, researchers have tried 600 other factors, but none were found to be significant. Among them, 500 are truly insignificant, but the other 100 are true factors. The total number of tests (M) is 700. Two types of mistakes are made in this process: 50 factors are falsely discovered to be true (type I error or false positive), while 100 true factors are buried in unpublished work (type II error or false negative). The usual statistical control in a multiple testing context aims at reducing "50" or "50/100," the absolute or proportionate occurrence of

[9] For early research on multiple testing, see Tukey (1951, 1953) for Tukey's range test and Scheffé (1959) for Scheffé's method on adjusting significance levels in a multiple regression context.

[10] See Shaffer (1995) for a review of multiple testing procedures that control for the family-wise error rate. See Farcomeni (2007) for a review that focuses on procedures that control the false-discovery rate.

[11] See Shanken (1990), Ferson and Harvey (1999), and Boudoukh et al. (2007).

false discoveries, respectively. Of course, we only observe published factors because factors that are tried and found to be insignificant rarely make it to publication.[12] This poses a challenge since the usual statistical techniques only handle the case in which all test results are observable.

Panel B defines the corresponding terms in a formal statistical testing framework. In a factor testing exercise, the typical null hypothesis is that a factor is not significant. Therefore, a factor being insignificant means the null hypothesis is "true." Using "0" ("1") to indicate the null is true (false) and "a" ("r") to indicate "not reject" ("reject"), we can easily summarize panel A. For instance, $N_{0|r}$ measures the number of rejections when the null is true (i.e., the number of false discoveries) and $N_{1|a}$ measures the number of failed rejections when the null is false (i.e. the number of missed discoveries). To avoid confusion, we try not to use standard statistical language in describing our notation but rather use words unique to our factor testing context. The generic notation in panel B is convenient in formally defining different types of errors and describing adjustment procedures in subsequent sections.

3.3 Type I and type II errors

For a single hypothesis test, a value α is used to control type I error rate: the probability of finding a factor to be significant when it is not. The α is sometimes called the "level of significance." In a multiple testing framework, restricting each individual test's type I error rate at α is not enough to control the overall probability of false discoveries. The intuition is that, under the null that all factors are insignificant, it is very likely for an event with α probability to occur when many factors are tested. In multiple hypothesis testing, we need measures of the type I error that help us simultaneously evaluate the outcomes of many individual tests.

To gain some intuition about plausible measures of type I error, we return to panel B of Table 2. $N_{0|r}$ and $N_{1|a}$ count the total number of the two types of errors: $N_{0|r}$ counts false discoveries, while $N_{1|a}$ counts missed discoveries. As generalized from single hypothesis testing, the type I error in multiple hypothesis testing is also related to false discoveries, by which we conclude a factor is "significant" when it is not. But, by definition, we must draw several conclusions in multiple hypothesis testing, and there is a possible false discovery for each. Therefore, plausible definitions of the type I error should take into account the joint occurrence of false discoveries.

The literature has adopted at least two ways of summarizing the "joint occurrence." One approach is to count the total number of false discoveries

---

[12] Examples of the publication of unsuccessful factors include Fama and MacBeth (1973) and Ferson and Harvey (1993). Fama and MacBeth (1973) show that squared beta and standard deviation of the market model residual have an insignificant role in explaining the cross-section of expected returns. However, the inclusion of these two variables was a result of a falsification experiment rather than a search for new factors. Overall, it is rare to publish "nonresults" and all instances of published nonresults are coupled with significant results for other factors.

$N_{0|r}$. $N_{0|r}$ greater than zero suggests incorrect statistical inference for the overall multiple testing problem—the occurrence of which we should limit. Therefore, the probability of event $N_{0|r} > 0$ should be a meaningful quantity for us to control. Indeed, this is the intuition behind the family-wise error rate introduced later. On the other hand, when the total number of discoveries R is large, one or even a few false discoveries may be tolerable. In this case, $N_{0|r}$ is no longer a suitable measure; a certain false discovery proportion may be more desirable. Unsurprisingly, the expected value of $N_{0|r}/R$ is the focus of false discovery rate, the second type of control.

3.3.1 Family-wise error rate. The two aforementioned measures are the most widely used in the statistics literature. Moreover, many other techniques can be viewed as extensions of these measures. Holm (1979) is the first to formally define the family-wise error rate. Benjamini and Hochberg (1995) define and study the false discovery rate. Alternative definitions of error rate include per comparison error rate (Saville 1990), positive false discovery rate (Storey 2003), and generalized false discovery rate (Sarkar and Guo 2009). We now describe the two leading approaches in detail.

The family-wise error rate (FWER) is the probability of at least one type I error:

$$\text{FWER} = Pr(N_{0|r} \geq 1).$$

FWER measures the probability of even a single false discovery, regardless of the total number of tests. For instance, researchers might test 100 factors; FWER measures the probability of incorrectly identifying one or more factors to be significant. Given significance or threshold level α, we explore two existing methods (Bonferroni and Holm's adjustment) to ensure FWER does not exceed α. Even as the number of trials increases, FWER still measures the probability of at least one false discovery. This absolute control is in contrast to the proportionate control afforded by the false discovery rate (FDR), defined below.

3.3.2 False discovery rate. The *false discovery proportion* (FDP) is the proportion of type I errors:

$$\text{FDP} = \begin{cases} \dfrac{N_{0|r}}{R} & \text{if } R > 0, \\ 0 & \text{if } R = 0. \end{cases}$$

The false discovery rate (FDR) is defined as

$$\text{FDR} = E[FDP].$$

FDR measures the expected proportion of false discoveries among all discoveries. It is less stringent (i.e., leads to more discoveries) than FWER and usually much less so when many tests are performed.[13] Intuitively, this is because FDR allows $N_{0|r}$ to grow in proportion to R, whereas FWER measures the probability of making even a single type I error.

Returning to example A, panel A shows that a false discovery event has occurred under FWER since $N_{0|r}=50\geq 1$ and the realized FDP is high, $50/100=50\%$. This suggests that the probability of false discoveries (FWER) and the expected proportion of false discoveries (FDR) may both be high.[14] The remedy, as suggested by many FWER and FDR adjustment procedures, is to lower p-value thresholds for these hypotheses (p-value, as defined in our context, is the single test probability of having a t-statistic that is at least as large as the observed one under the null hypothesis). In terms of panel A, this would turn some of the fifty false discoveries insignificant and push them into the "Unpublished" category. Hopefully, the fifty true discoveries would survive this change in p-value threshold and remain significant, which is only possible if their p-values are relatively small.

On the other hand, type II errors—the mistake of missing true factors—are also important in multiple hypothesis testing. Similar to type I errors, both the total number of missed discoveries $N_{1|a}$ and the fraction of missed discoveries among all abandoned tests $N_{1|a}/(M-R)$ are frequently used to measure the severity of type II errors.[15] Ideally, one would like to simultaneously minimize the chance of committing a type I error and that of committing a type II error. In our context, we would like to include as few insignificant factors (i.e., as low a type I error rate) as possible and simultaneously as many significant ones (i.e., as low a type II error rate) as possible. Unfortunately, this is not

---

13 There is a natural ordering between FDR and FWER. Theoretically, FDR is always bounded above by FWER, that is, $FDR\leq FWER$. To see this, by definition,

$$\begin{aligned}\text{FDR} &= E[\frac{N_{0|r}}{R}|R>0]Pr(R>0)\\ &\leq E[I_{(N_{0|r}\geq 1)}|R>0]Pr(R>0)\\ &= Pr((N_{0|r}\geq 1)\cap(R>0))\\ &\leq Pr(N_{0|r}\geq 1)=\text{FWER},\end{aligned}$$

where $I_{(N_{0|r}\geq 1)}$ is an indicator function of event $N_{0|r}\geq 1$. This implies that procedures that control FWER under a certain significance level automatically control FDR under the same significance level. In our context, a factor discovery criterion that controls FWER at α also controls FDR at α.

14 Panel A only shows one realization of the testing outcome for a certain testing procedure (e.g., single tests). To evaluate FWER and FDR, both of which are expectations and hence depend on the underlying joint distribution of the testing statistics, we need to know the population of the testing outcomes. To give an example that is compatible with panel A, we assume that the t-statistics for the 700 hypotheses are independent. Moreover, we assume the t-statistic for a false factor follows a normal distribution with mean of zero and variance of 1.0, that is, $\mathcal{N}(0,1)$; for a true factor, we assume its t-statistic follows a normal distribution with mean of 2.0 and variance of 1.0, that is, $\mathcal{N}(2,1)$. Under these assumptions about the joint distribution of the test statistics, we find via simulations that FWER is 100% and FDR is 26%, both exceeding 5%.

15 See Simes (1986) for one example of type II error in simulation studies and Farcomeni (2007) for another example in medical experiments.

feasible: as in single hypothesis testing, a decrease in the type I error rate often leads to an increase in the type II error rate and vice versa. We therefore seek a balance between the two types of errors. A standard approach is to specify a significance level α for the type I error rate and derive testing procedures that aim to minimize the type II error rate, that is, maximize power, among the class of tests with type I error rate at most α.

When comparing two testing procedures that can both achieve a significance level α, it seems reasonable to use their type II error rates. However, when we have multiple tests, the exact type II error rate typically depends on a set of unknown parameters and is therefore difficult to assess.[16] To overcome this difficulty, researchers frequently use the distance of the actual type I error rate to some prespecified significance level as the measure for a procedure's efficiency. Intuitively, if a procedure's actual type I error rate is strictly below α, we can probably push this error rate closer to α by making the testing procedure less stringent, that is, a higher p-value threshold so there will be more discoveries. In doing so, the type II error rate is presumably lowered given the inverse relation between the two types of error rates. Therefore, once a procedure's actual type I error rate falls below a prespecified significance level, we want it to be as close as possible to that significance level in order to achieve the smallest type II error rate. Ideally, we would like a procedure's actual type I error rate to be exactly the same as the given significance level.[17]

Both FWER and FDR are important concepts that are widely applied in many scientific fields. However, based on specific applications, one may be preferred over the other. When the number of tests is very large (e.g., a million), FWER controlling procedures tend to become very tough as they control for the occurrence of even a single false discovery among one million tests. As a result, they often lead to a very limited number of discoveries, if any. Conversely, FWER control is more desirable when the number of tests is relatively small, in which case more discoveries can be achieved and at the same time trusted. In the context of our paper, we are sure that many tests have been tried in the finance literature. Although there are around 300 published ones, hundreds or even thousands of factors could have been constructed and tested. However, it is not clear whether this number should be considered "large" compared to the

---

16 In single hypothesis testing, the type II error is a function of the unknown true parameter value—in our context, the population factor mean return—under the alternative hypothesis. By tracing out all possible values under the alternative hypothesis, we obtain the type II error function. The situation is more complicated in multiple hypothesis testing because the type II error depends on multiple parameters that correspond to the collection of alternative hypotheses for all the tests. Hence, the type II error function is multivariate when there are multiple tests. See Zehetmayer and Posch (2010) for power estimation in large-scale multiple testing problems.

17 In our framework, individual p-values are sufficient statistics for us to make adjustment for multiple tests. Each individual p-value represents the probability of having a t-statistic that is at least as large as the observed one under the null hypothesis. What happens under the alternative hypotheses (i.e., type II error rate) does not directly come into play because hypothesis testing in the frequentist framework has a primary focus on the type I error rate. When we deviate from the frequentist framework and consider Bayesian methods, the type II error rate becomes more important because Bayesian odds ratios put the type I and type II error rates on the same footing.

Table 3
A summary of *p*-value adjustments

Adjustment type	Single/Sequential	Multiple test
Bonferroni	single	family-wise error rate
Holm	sequential	family-wise error rate
Benjamini, Hochberg, and Yekutieli (BHY)	sequential	false discovery rate

number of tests conducted in, say, medical research.[18] This creates difficulty in choosing between FWER and FDR. Given this difficulty, we do not take a stand on the relative appropriateness of these two measures but instead provide adjusted p-values for both. Researchers can compare their p-values with these benchmarks to see whether FDR or even FWER is satisfied.

3.4 *p*-value adjustment: Three approaches

The statistics literature has developed many methods to control both FWER and FDR.[19] We choose to present the three most well-known adjustments: Bonferroni, Holm, and Benjamini, Hochberg, and Yekutieli (BHY). Both Bonferroni and Holm control FWER, and BHY controls FDR. Depending on how the adjustment is implemented, they can be categorized into two general types of corrections: a "single-step" correction equally adjusts each p-value, and a "sequential" correction is an adaptive procedure that depends on the entire distribution of p-values. Bonferroni is a single-step procedure, whereas Holm and BHY are sequential procedures. Table 3 summarizes the two properties of the three methods.

Suppose there are in total M tests and we choose to set FWER at α_w and FDR at α_d. In particular, we consider an example with the total number of tests $M=10$ to illustrate how different adjustment procedures work. For our main results, we set both α_w and α_d at 5%. Table 4, panel A, lists the t-statistics and the corresponding p-values for ten hypothetical tests. The numbers in the table are broadly consistent with the magnitude of t-statistics that researchers report for factor significance. Note that all ten factors will be "discovered" if we test one hypothesis at a time. Multiple testing adjustments will usually generate different results.[20]

3.4.1 Bonferroni's adjustment. Bonferroni's adjustment is as follows:

- Reject any hypothesis with p-value $\leq \frac{\alpha_w}{M}$:

$$p_i^{Bonferroni} = \min[M \times p_i, 1].$$

[18] For instance, tens of thousands of tests are performed in the analysis of DNA microarrays. See Farcomeni (2007) for more applications of multiple testing in medical research.

[19] Methods that control FWER include Holm (1979), Hochberg (1988), and Hommel (1988). Methods that control FDR include those of Benjamini and Hochberg (1995), Benjamini and Liu (1999), and Benjamini and Yekutieli (2001).

[20] Readers who are already familiar with the three multiple testing adjustment procedures can skip to Section 4.5 for our main results.

Table 4
An example of multiple testing

Panel A: Single tests and "significant" factors

Test →	1	2	3	4	5	6	7	8	9	10	# of discoveries
t-statistic	1.99	2.63	2.21	3.43	2.17	2.64	4.56	5.34	2.75	2.49	10
p-value (%)	**4.66**	**0.85**	**2.71**	**0.05**	**3.00**	**0.84**	**0.00**	**0.00**	**0.60**	**1.28**	

Panel B: Bonferroni "significant" factors

Test →	1	2	3	4	5	6	7	8	9	10	
t-statistic	1.99	2.63	2.21	3.43	2.17	2.64	4.56	5.34	2.75	2.49	3
p-value (%)	4.66	0.85	2.71	**0.05**	3.00	0.84	**0.00**	**0.00**	0.60	1.28	

Panel C: Holm adjusted p-values and "significant" factors

Reordered tests b	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	
Old order	8	7	4	9	6	2	10	3	5	1	4
p-value (%)	**0.00**	**0.00**	**0.05**	**0.60**	0.84	0.85	1.28	2.71	3.00	4.66	
$\alpha_w/(M+1-b)$ $\alpha_w=5\%$	0.50	0.56	0.63	0.71	0.83	1.00	1.25	1.67	2.50	5.00	

Panel D: BHY adjusted p-values and "significant" factors

Reordered tests b	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	
Old order	8	7	4	9	6	2	10	3	5	1	6
p-value (%)	**0.00**	**0.00**	**0.05**	**0.60**	**0.84**	**0.85**	1.28	2.71	3.00	4.66	
$(b\cdot\alpha_d)/(M\times c(M))$ $\alpha_d=5\%$	0.17	0.34	0.51	0.68	0.85	1.02	1.19	1.37	1.54	1.71	

The table displays ten t-statistics and their associated p-values for a hypothetical example. Panels A and B highlight the significant factors under single tests and Bonferroni's procedure, respectively. Panels C and D explain Holm's and BHY's adjustment procedure, respectively. The bold numbers in each panel are associated with significant factors under the specific adjustment procedure of that panel. M represents the total number of tests (M = 10) and $c(M)=\sum_{j=1}^{M} 1/j$. b is the order of p-values from lowest to highest. α_w is the significance level for Bonferroni's and Holm's procedure, and α_d is the significance level for BHY's procedure. Both numbers are set at 5%. The cutoff p-value for Bonferroni is 0.5%, for Holm is 0.60%, and for BHY is 0.85%.

Bonferroni applies the same adjustment to each test. It inflates the original p-value by the number of tests M; the adjusted p-value is compared with the threshold value α_w.

Example 4.4.1 To apply Bonferroni's adjustment to the example in Table 4, we simply multiply all the p-values by ten and compare the new p-values with $\alpha_w=5\%$. Equivalently, we can look at the original p-values and consider the cutoff of $0.5\%(=\alpha_w/10)$. This leaves the t-statistic of tests 4, 7, and 8 as significant, which are highlighted in panel B.

Using the notation in panel B of Table 2 and assuming M_0 of the M null hypotheses are true, Bonferroni operates as a single-step procedure that can be shown to restrict FWER at levels less than or equal to $(M_0\times\alpha_w)/M$, without any assumption on the dependence structure of the p-values. Since $M_0\leq M$, Bonferroni also controls FWER at level α_w.[21]

[21] The number of true nulls M_0 is unknown, so we usually cannot make Bonferroni more powerful by increasing α_w to $\hat{\alpha}=M\alpha_w/M_0$ (note that $M_0\hat{\alpha}/M=\alpha_w$). However some papers, including those by Schweder and Spjotvoll

3.4.2 Holm's adjustment. Sequential methods have been proposed to adjust p-values in multiple hypothesis testing.[22] They are motivated by a seminal paper by Schweder and Spjotvoll (1982), who suggest a graphical presentation of the multiple testing p-values. In particular, using N_p to denote the number of tests that have a p-value exceeding p, Schweder and Spjotvoll (1982) suggest plotting N_p against $(1-p)$. When p is not very small (e.g., $p > 0.2$), it is very likely that the associated test is from the null hypothesis. In this case, the p-value for a null test can be shown to be uniformly distributed between 0 and 1. It then follows that for a large p and under independence among tests, the expected number of tests with a p-value exceeding p equals $T_0(1-p)$, where T_0 is the number of null hypotheses, i.e., $E(N_p) = T_0(1-p)$. By plotting N_p against $(1-p)$, the graph should be approximately linear with slope T_0 for large p-values. Points on the graph that substantially deviate from this linear pattern should correspond to non-null hypotheses, i.e., discoveries. The gist of this argument—large and small p-values should be treated differently—has been distilled into many variations of sequential adjustment methods, among which we will introduce Holm's method that controls FWER and BHY's method that controls FDR.

Holm's adjustment is as follows:

- Order the original p-values such that $p_{(1)} \leq p_{(2)} \leq \cdots p_{(b)} \leq \cdots \leq p_{(M)}$, and let the associated null hypotheses be $H_{(1)}, H_{(2)}, \cdots H_{(b)} \cdots, H_{(M)}$.
- Let k be the minimum index such that $p_{(b)} > \frac{\alpha_w}{M+1-b}$.
- Reject the null hypotheses $H_{(1)} \cdots H_{(k-1)}$ (i.e., declare these factors significant), but not $H_{(k)} \cdots H_{(M)}$.

The equivalent adjusted p-value is therefore

$$p_{(i)}^{Holm} = \min[\max_{j \leq i}\{(M-j+1)p_{(j)}\}, 1].$$

Holm's adjustment is a step-down procedure: for the ordered p-values, we start from the smallest p-value and go down to the largest one.[23] If k is the smallest index that satisfies $p_{(b)} > \frac{\alpha_w}{M+1-b}$, we will reject all tests whose ordered index is below k.

To explore how Holm's adjustment procedure works, suppose k is the smallest index such that $p_{(b)} > \frac{\alpha_w}{M+1-b}$. This means that for $b < k$, $p_{(b)} \leq \frac{\alpha_w}{M+1-b}$. In particular, for $b = 1$, Bonferroni equals Holm, that is, $\frac{\alpha_w}{M} = \frac{\alpha_w}{M+1-(b=1)}$; for $b = 2$,

(1982) and Hochberg and Benjamini (1990), try to improve the power of Bonferroni by estimating M_0. We try to achieve the same goal by using either Holm's procedure, which also controls FWER, or procedures that control FDR, an alternative definition of type I error rate.

22 Here, "sequential" refers to the fact that we adjust the ordered p-values sequentially. It does not mean that the individual tests arrive sequentially.

23 Viewing small p-values as "up" and large p-values as "down," Holm's procedure is a "step-down" procedure in that it goes from small p-values to large ones. This terminology is consistent with the statistics literature. Of course, small p-values are associated with "large" values of the test statistics.

$\frac{\alpha_w}{M} < \frac{\alpha_w}{M+1-(b=2)}$, so Holm is less stringent than Bonferroni. Since less stringent hurdles are applied to the second to the $(k-1)$th p-values, more discoveries are generated under Holm's than Bonferroni's adjustment.

Example 4.4.2 To apply Holm's adjustment to the example in Table 4, we first order the p-values in ascending order and try to locate the smallest index k that makes $p_{(b)} > \frac{\alpha_w}{M+1-b}$. Table 4, panel C, shows the ordered p-values and the associated $\frac{\alpha_w}{M+1-b}$'s. Starting from the smallest p-value and going up, we see that $p_{(b)}$ is below $\frac{\alpha_w}{M+1-b}$ until $b=5$, at which point $p_{(5)}$ is above $\frac{\alpha_w}{10+1-5}$. Therefore, the smallest b that satisfies $p_{(b)} > \frac{\alpha_w}{M+1-b}$ is 5 and we reject the null hypothesis for the first four ordered tests (we discover four factors) and fail to reject the null for the remaining six tests. The original labels for the rejected tests are in the second row in panel C. Compared to Bonferroni, one more factor (test 9) is discovered; that is, four factors rather than three are significant. In general, Holm's approach leads to more discoveries and all discoveries under Bonferroni are also discoveries under Holm's criteria.

Like Bonferroni, Holm also restricts FWER at α_w without any requirement on the dependence structure of p-values. It can also be shown that Holm is uniformly more powerful than Bonferroni in that tests rejected (factors discovered) under Bonferroni are always rejected under Holm, but not vice versa. In other words, Holm leads to at least as many discoveries as Bonferroni. Given the dominance of Holm over Bonferroni, one might opt to only use Holm. We include Bonferroni because it is the most widely used adjustment and a simple single-step procedure.

3.4.3 Benjamini, Hochberg, and Yekutieli's adjustment. Benjamini, Hochberg, and Yekutieli's (BHY) adjustment is as follows:

- As with Holm's procedure, we order the original p-values such that $p_{(1)} \leq p_{(2)} \leq \cdots p_{(b)} \leq \cdots \leq p_{(M)}$ and let associated null hypotheses be $H_{(1)}, H_{(2)}, \cdots H_{(b)} \cdots, H_{(M)}$.
- Let k be the maximum index such that $p_{(b)} \leq \frac{b}{M \times c(M)} \alpha_d$.
- Reject null hypotheses $H_{(1)} \cdots H_{(k)}$, but not $H_{(k+1)} \cdots H_{(M)}$.

The equivalent adjusted p-value is defined sequentially as

$$p_{(i)}^{BHY} = \begin{cases} p_{(M)} & \text{if } i = M, \\ \min[p_{(i+1)}^{BHY}, \frac{M \times c(M)}{i} p_{(i)}] & \text{if } i \leq M-1, \end{cases}$$

where, $c(M)$ is a function of the total number of tests M and controls for the generality of the test. The larger $c(M)$ is the more stringent test and hence is more general in guarding against dependency among the test statistics. In particular, Benjamini and Yekutieli (2001) show that setting $c(M)$ at

$$c(M) = \sum_{j=1}^{M} \frac{1}{j} \tag{1}$$

allows the procedure to work under arbitrary dependency among the test statistics. We focus on this specification due to its generality but will discuss what happens under alternative specifications of $c(M)$.

In contrast to Holm's, BHY's method starts with the largest p-value and goes to the smallest one. If k is the largest index that satisfies $p_{(b)} \leq \frac{b}{M \times c(M)}\alpha_d$, we will reject tests (discover factors) whose ordered index is below or equal to k. Also, note that α_d (significance level for FDR) is chosen to be the same as α_w (significance level for FWER). The significance level is subjective in nature. Here, we choose the same significance level to make an apples-to-apples comparison between FDR and FWER adjustment procedures. We discuss this choice in more detail in Section 4.6.

To explore how BHY works, let k be the largest index, such that $p_{(b)} \leq \frac{b}{M \times c(M)}\alpha_d$. This means that for $b > k$, $p_{(b)} > \frac{b}{M \times c(M)}\alpha_d$. In particular, we have $p_{(k+1)} > \frac{(k+1)}{M \times c(M)}\alpha_d$, $p_{(k+2)} > \frac{(k+2)}{M \times c(M)}\alpha_d$, ..., $p_{(M)} > \frac{M}{M \times c(M)}\alpha_d$. We see that the $(k+1)$th to the last null hypotheses, not rejected, are compared to numbers smaller than α_d, the usual significance level in single hypothesis testing. By being stricter than single hypothesis tests, BHY guarantees that the false discovery rate, which depends on the joint distribution of all the test statistics, is below the prespecified significance level. See Benjamini and Yekutieli (2001) for details on the proof.

Example 4.4.3 To apply BHY's adjustment to the example in Table 4, we first order the p-values in ascending order and try to locate the largest index k that satisfies $p_{(b)} \leq \frac{b}{M \times c(M)}\alpha_d$. Table 4, panel D, shows the ordered p-values and the associated $\frac{b}{M \times c(M)}\alpha_d$'s. Starting from the largest p-value and going down, we see that $p_{(b)}$ is above $\frac{b}{M \times c(M)}\alpha_d$ until $b=6$, at which point $p_{(6)}$ is below $\frac{6}{10 \times 2.93}\alpha_d$. Therefore, the smallest b that satisfies $p_{(b)} \leq \frac{b}{M \times c(M)}\alpha_d$ is 6, and we reject the null hypothesis for the first six ordered tests and fail to reject for the remaining four tests. In the end, BHY leads to six significant factors (tests 8, 7, 4, 9, 6, and 2), three more than Bonferroni and two more than Holm.

In summary, for single tests, using the usual 5% cutoff, 10 out of 10 are discovered. Allowing for multiple tests, the cutoffs are far smaller, with BHY at 0.85%, Holm at 0.60%, and Bonferroni at 0.5%.

The choice of $c(M)$ determines the generality of BHY's procedure. Intuitively, when $c(M)$ is larger, then the more difficult it is to satisfy the inequality $p_{(b)} \leq \frac{b}{M \times c(M)}\alpha_d$, and hence there will be fewer discoveries. This makes it easier to restrict the false discovery rate below a given significance level since fewer discoveries are made. In the original work that develops the concept of false discovery rate and related testing procedures, $c(M)$ is set equal to one. Under this choice, BHY is only valid when the test statistics are independent or positively dependent. With our choice of $c(M)$ (i.e., $c(M) = \sum_{j=1}^{M} \frac{1}{j}$), BHY is valid under any form of dependence among the

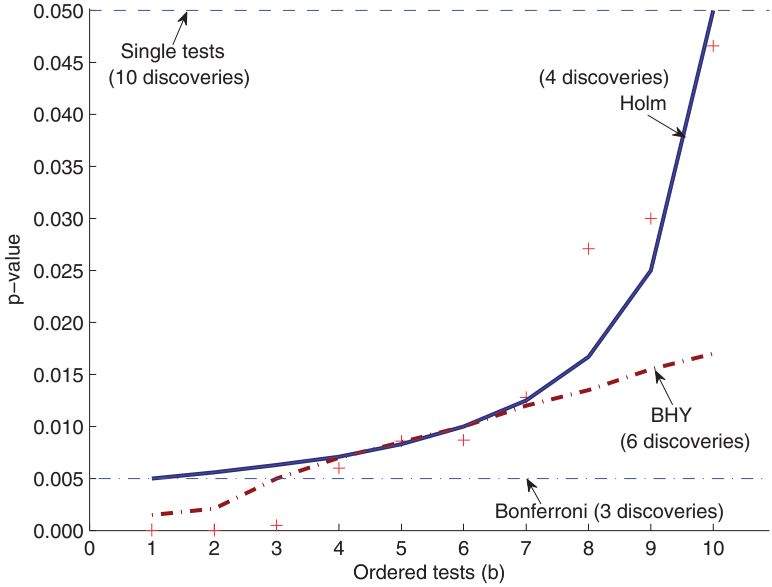

Figure 1
Multiple test thresholds for example A,
The ten p-values for the example in Table 4 and the adjusted p-value lines for various adjustment procedures. All ten factors are discovered using the standard criteria for single tests, three under Bonferroni, four under Holm, and six under BHY. The significance level is set at 5% for each adjustment method.

p-values.[24] Note with $c(M) > 1$, this reduces the size of $\frac{b}{M \times c(M)}\alpha_d$ and it is tougher to satisfy the inequality $p_{(b)} \leq \frac{b}{M \times c(M)}\alpha_d$. That is, there will be fewer factors found to be significant.

Figure 1 summarizes our example. It plots the original p-values (single tests), as well as adjusted p-value lines, for various multiple testing adjustment procedures. We see the stark difference in outcomes between multiple and single hypothesis testing. While all ten factors would be discovered under single hypothesis testing, only three to six factors would be discovered under a multiple hypothesis test. Although single hypothesis testing guarantees the type I error of each test meets a given significance level, meeting the more stringent FWER or FDR bound will lead us to discard a number of factors.

3.5 Summary statistics

Figure 2 shows the history of discovered factors and publications.[25] We observe a dramatic increase in factor discoveries during the last decade. In the early

[24] See Benjamini and Yekutieli (2001) for the proof.

[25] To be clear, we only count those that have t-statistics or equivalent statistics reported. Roughly twenty new factors fail to satisfy this requirement. For additional details, see factors in Table 6 marked with ‡.

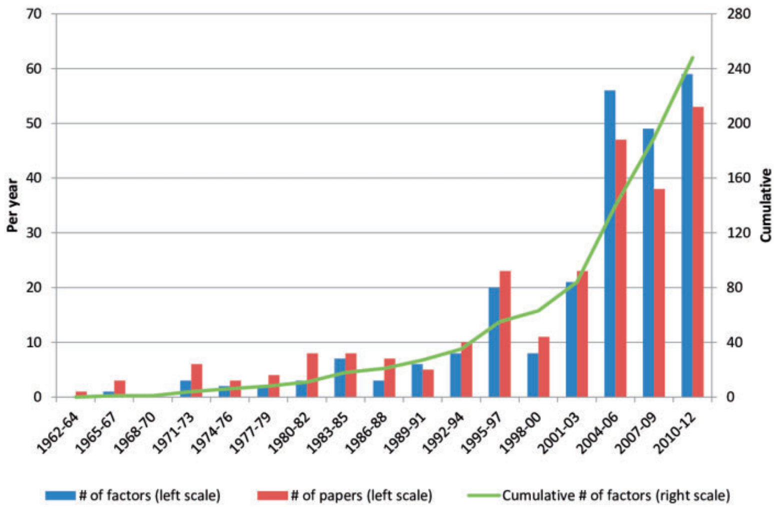

Figure 2
Factors and publications.

period from 1980 to 1991, only about one factor is discovered per year. This number has grown to around five for the 1991–2003 period, during which time a number of papers, such as Fama and French (1992), Carhart (1997), and Pastor and Stambaugh (2003), spurred interest in studying cross-sectional return patterns. In the last nine years, the annual factor discovery rate has increased sharply to around 18. In total, 164 factors were discovered in the past nine years, roughly doubling the 84 factors discovered in all previous years. We do not include working papers in Figure 2. In our sample, there are 63 working papers covering 68 factors.

We obtain t-statistics for each of the 316 factors discovered, including the ones in the working papers.[26] The overwhelming majority of t-statistics exceed the 1.96 benchmark for 5% significance.[27] The nonsignificant ones typically belong to papers that propose a number of factors. These likely represent only a small subsample of nonsignificant t-statistics for all tried factors. Importantly, we take published t-statistics as given. That is, we assume they are econometrically sound with respect to the usual suspects (data errors, coding errors, misalignment, heteroscedasticity, autocorrelation, clustering, outliers, etc.).

---

[26] The sign of a t-statistic depends on the direction of the long/short strategy. We usually calculate p-values based on two-sided t-tests, so the sign does not matter. From an investment perspective, the sign of the mean return of a long/short strategy does not matter as we can always reverse the direction of the strategy. Therefore we use absolute values of these t-statistics.

[27] The multiple testing framework is robust to outliers. The procedures are based on either the total number of tests (Bonferroni) or the order statistics of t-statistics (Holm and BHY).

3.6 p-value adjustment when all tests are published ($M = R$)

We now apply the three adjustment methods previously introduced to the observed factor tests, under the assumption that the test results of all tried factors are available. We know that this assumption is false since our sample underrepresents all insignificant factors by conventional significance standards: we only observe those insignificant factors that are the results of purposeful falsification experiments. We design methods to handle this missing data issue later.

Despite some limitations, our results in this section are useful for at least two reasons. First, the benchmark t-statistic based on our incomplete sample provides a lower bound of the true t-statistic benchmark. In other words, if M (total number of tests) $>$ R (total number of discoveries), then we would expect fewer factors than when $M = R$,[28] so future t-statistics need to at least surpass our benchmark to claim significance. Second, results in this section can be rationalized within a Bayesian or hierarchical testing framework.[29] Factors in our list constitute an "elite" group: they have survived academia's scrutiny for publication. Placing a high prior on this group in a Bayesian testing framework or viewing this group as a cluster in a hierarchical testing framework, one can interpret results in this section as the first-step factor selection within an a priori group.

Based on our sample of observed t-statistics of published factors,[30] we obtain three benchmark t-statistics. In particular, at each point in time, we transform the set of available t-statistics into p-values. We then apply the three adjustment methods to obtain benchmark p-values. Finally, these p-value benchmarks are transformed back into t-statistics, assuming that standard normal distribution approximates the t-distribution well. To guide future research, we extrapolate our benchmark t-statistics into the future, assuming that the rate of "factor production" remains the same as the recent history, that is, 2003–2012.

We choose to set α_w at 5% (Holm, FWER) and α_d at 1% (BHY, FDR) for our main results. The significance level is subjective, as in individual hypothesis testing, where conventional significance levels are usually adopted. Since FWER is a special case of the type I error in individual testing and 5% seems the default significance level in cross-sectional studies, we set α_w at 5%. On the other hand, FDR is a more lenient control relative to FWER. If we choose the same α_d as α_w, then by definition the BHY method will be more lenient than both Holm and Bonferroni. We set FDR at 1% but will explain what happens when α_d is increased to 5%.

Figure 3 presents the three benchmark t-statistics. Both Bonferroni and Holm adjusted benchmark t-statistics are monotonically increasing in the number of discoveries. For Bonferroni, the benchmark t-statistic starts at 1.96 and increases to 3.78 by 2012. It reaches 4.00 in 2032. The corresponding p-values

28 This is always true for Bonferroni's adjustment but is not always true for the other two types of adjustments. The Bonferroni adjusted t-statistic is monotonically increasing in the number of trials, so the t-statistic benchmark

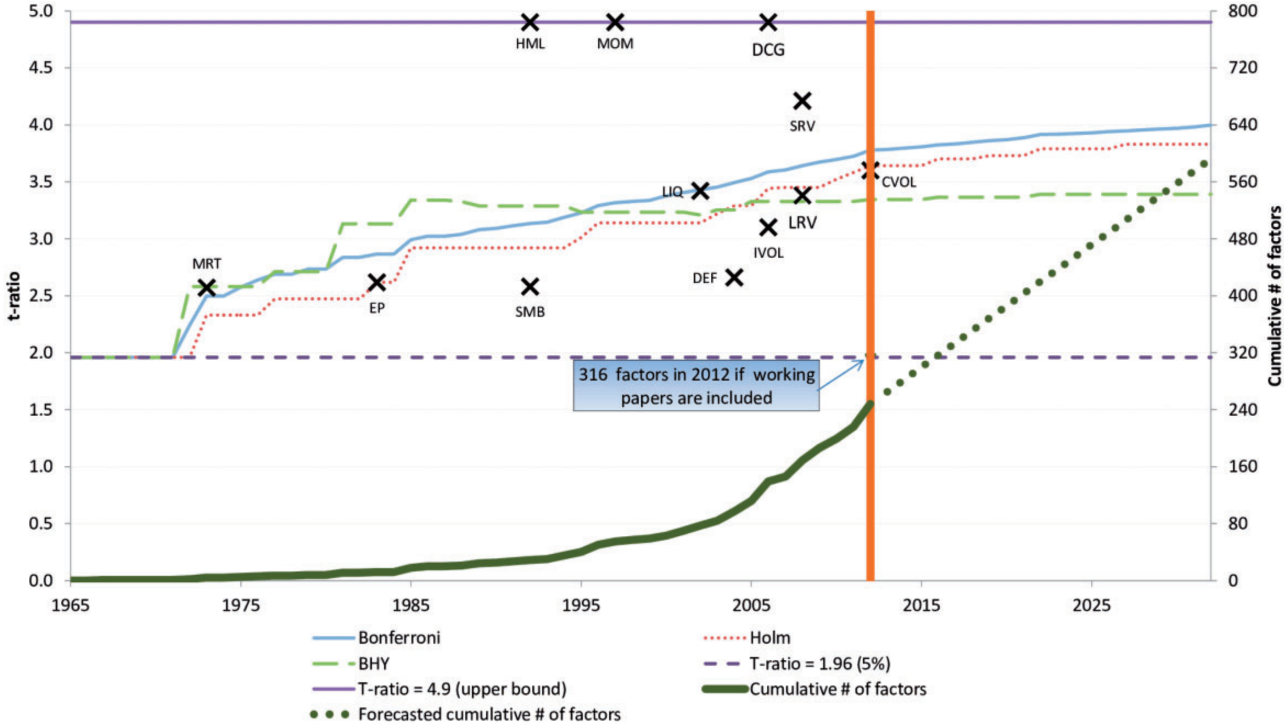

Figure 3
Adjusted *t*-statistics, 1965–2032.
Bonferroni and Holm are multiple testing adjustment procedures that control the family-wise error rate (FWER) and are described in Sections 4.4.1 and 4.4.2, respectively. BHY is a multiple testing adjustment procedure that controls the false discovery rate (FDR) and is described in Section 4.4.3. The green solid curve shows the historical cumulative number of factors discovered, excluding those from working papers. Forecasts (dotted green line) are based on a linear extrapolation. The dark crosses mark selected factors proposed by the literature. They are MRT (market beta; Fama and MacBeth 1973), EP (earnings-price ratio; Basu 1983), SMB and HML (size and book-to-market; Fama and French (1992)), MOM (momentum; Carhart 1997), LIQ (liquidity; Pastor and Stambaugh 2003), DEF (default likelihood; Vassalou and Xing 2004), IVOL (idiosyncratic volatility; Ang et al. 2006); DCG (durable consumption goods; Yogo 2006), SRV and LRV (short-run and long-run volatility; Adrian and Rosenberg 2008), and CVOL (consumption volatility; Boguth and Kuehn 2012). *t*-statistics over 4.9 are truncated at 4.9. For detailed descriptions of these factors, see Table 6 and http://faculty.fuqua.duke.edu/~charvey/Factor-List.xlsx.

(under single tests) for 3.78 and 4.00 are 0.02% and 0.01%, respectively, much lower than the starting level of 5%. Holm implied t-statistics always fall below Bonferroni t-statistics, consistent with the fact that Bonferroni always results in fewer discoveries than Holm. However, Holm tracks Bonferroni closely and their differences are small. BHY implied benchmarks, on the other hand, are not monotonic. They fluctuate before year 2000 and stabilize at 3.39 (p-value = 0.07%) after 2010. This stationarity feature of BHY implied t-statistics, inherent in the definition of FDR, contrasts with that of Bonferroni and Holm. Intuitively, at any fixed significance level α, the law of large numbers forces the false discovery rate (FDR) to converge to a constant. If we change α_d to 5%, the corresponding BHY implied benchmark t-statistic is 2.78 (p-value = 0.54%) in 2012 and 2.81 (p-value = 0.50%) in 2032, still much higher than the starting value of 1.96. In sum, after taking testing multiplicity into account, we believe the minimum threshold t-statistic for 5% significance is about 2.8, which corresponds to a p-value (if a single test) of 0.5%.

To see how the new t-statistic benchmarks better reveal the statistical significance of factors, in Figure 3 we mark the t-statistics of a few prominent factors. Among these factors, HML, MOM, DCG, SRV, and MRT are significant across all types of t-statistic adjustments, EP, LIQ, and CVOL are sometimes significant, and the rest are never significant.

One concern with our results is that factors are discovered at different times and tests are conducted using different methods. This heterogeneity in the time of discovery and testing methods may blur the interpretation of our results. Ideally, we want updated factor tests that are based on the most recent sample and the same testing method.[31] To alleviate this concern, we focus on the group of factors that are published no earlier than 2000 and rely on Fama-MacBeth tests. Additionally, we require that factor tests cover at least the 1970–1995 period and have as controls at least the Fama-French three factors (Fama and French 1993). This leaves us with 124 factors. Based on this factor group, the Bonferroni and Holm implied threshold t-statistics are 3.54 and 3.20 (5% significance), respectively, and the BHY implied thresholds are 3.23 (1% significance) and 2.67 (5% significance) by 2012. Not surprisingly, these statistics are smaller than the corresponding thresholds based on the full sample.

will only rise if there are more factors. Holm and BHY depend on the exact t-statistic distribution, so more factors do not necessarily imply a higher t-statistic benchmark.

[29] See Wagenmakers and Grünwald (2006) and Storey (2003) on Bayesian interpretations of traditional hypothesis testing. See Meinshausen (2008) for a hierarchical approach on variable selection.

[30] See the Online Appendix B for details on our sampling procedure.

[31] We want to stress that the three types of adjustments in our paper are robust to the heterogeneity in the time of discovery and testing methods among individual studies. That is, despite the varying degrees of sample overlap and the differences in the testing methods, our adjustment procedures guarantee that the type I errors (however they are defined) are controlled under their prespecified levels. Therefore, from a technical point of view, neither nonsimultaneity nor differences in testing methods invalidate our results.

However, the general message that we need a much higher t-statistic threshold when multiple testing is taken into account is unchanged.

3.7 Robustness

3.7.1 Test statistics dependence. There is a caveat for all three methods considered so far. In the context of multiple testing, any type of adjustment procedure can become too stringent when there is a certain dependence structure in the data. This is because these procedures are primarily designed to guard against type I errors. Under a certain correlation structure, they may penalize type I errors too harshly and lead to a high type II error rate.

In theory, under independence, Bonferroni and Holm approximately achieve the prespecified significance level α when the number of tests is large. On the other hand, both procedures tend to generate fewer discoveries than desired when there is a certain degree of dependence among the tests. Intuitively, in the extreme case in which all tests are the same (i.e., correlation = 1.0), we do not need to adjust at all: FWER is the same as the type I error rate for single tests. Hence, the usual single hypothesis test is sufficient. Similarly, BHY may generate too few discoveries when tests are independent or positively correlated.

Having discussed assumptions for the testing methods to work efficiently, we now try to think of scenarios that can potentially violate these assumptions. First, factors that proxy for the same type of risk may be dependent. Moreover, returns of long-short portfolios designed to achieve exposure to a particular type of factor may be correlated. For example, there are a number of factors with price in the denominator that are naturally correlated. We also count four different idiosyncratic volatility factors. If this type of positive dependence exists among test statistics, all three methods would likely generate fewer significant factors than desired. On the other hand, most often factors need to "stand their ground" to be published. In the end, if you think we are overcounting at 316, consider taking a haircut to 113 factors (the number of "common" factors in Table 1). Figure 3 shows that our main conclusions do not materially change. For example, the Holm at 113 factors is 3.29 (p-value = 0.10%), while Holm at 316 factors is 3.64 (p-value = 0.03%).

Second, research studying the same factor but based on different samples will generate highly dependent test statistics. Examples include the sequence of papers studying the size effect. We try to minimize this concern by including, with a few exceptions, only the original paper that proposes the factor. To the extent that our list includes few such duplicate factors, our method greatly reduces the dependence that would be introduced by including all papers studying the same factor but for different sample periods.

Finally, when dependence among test statistics can be captured by Pearson correlations among contemporaneous strategy returns, we present a new model in Section 5 to systematically incorporate the information in test correlations.

3.7.2 The case in which $M>R$. To deal with the hidden tests issue when $M>R$, we propose in Appendix A a simulation framework to estimate benchmark t-statistics. The idea is to first back out the underlying distribution for the t-statistics of all tried factors, then to generate benchmark t-statistic estimates, and apply the three adjustment procedures to simulated t-statistics samples.[32]

Based on our estimates, 71% of all tried factors are missing. Using this information, the new benchmark t-statistics for Bonferroni and Holm are estimated to be 4.01 and 3.96, respectively, both slightly higher than when $M=R$. This is as expected because more factors are tried under this framework. The BHY implied t-statistic increases from 3.39 to 3.68 at 1% significance and from 2.78 to 3.18 at 5% significance. In sum, across various scenarios, we think the minimum threshold t-statistic is 3.18, corresponding to BHY's adjustment for $M>R$ at 5% significance. Alternative cases all result in even higher benchmark t-statistics.

One concern with BHY is that our specification of $c(M)$ results in an overly stringent threshold for FDR. We therefore try the more lenient choice (i.e., $c(M)\equiv 1$) as in Benjamini and Hochberg (1995). Based on our estimate that 71% of tried factors are missing and by simulating the missing tests as in Appendix A, we find that the BHY implied threshold equals 3.05 at 5% significance and 3.17 at 1% significance. Indeed, these numbers are smaller than the numbers under our default specification of $c(M)$ (i.e., $c(M)=\sum_{j=1}^{M}\frac{1}{j}$). However, they are above 3.0 and therefore are consistent with our overall message.

3.7.3 A Bayesian hypothesis testing framework. We can also study multiple hypothesis testing within a Bayesian framework. One major obstacle of applying Bayesian methods in our context is the unobservability of all tried factors. While we propose new frequentist methods to handle this missing data problem, it is not clear how to structure the Bayesian framework in this context. In addition, the high dimensionality of the problem raises concerns on both the accuracy and the computational burden of Bayesian methods.

Nevertheless, ignoring the missing data issue, we outline a standard Bayesian multiple hypothesis testing framework in Appendix B and explain how it relates to our multiple testing framework. We discuss in detail the pros and cons of the Bayesian approach. In contrast to the frequentist approach, which uses generalized type I error rates to guide multiple testing, the Bayesian approach relies on the posterior likelihood function and thus contains a natural penalty term for multiplicity. However, this simplicity comes at the expense of having a restrictive hierarchical model structure and independence assumptions

32 The underlying assumption for the model in Appendix A is the independence among t-statistics, which may not be plausible given our previous discussions on test dependence. In that case, our structural model in Section 5 proposes a more realistic data generating process for the cross-section of test statistics.

that may not be realistic for our factor testing problem. Although extensions incorporating certain forms of dependence are possible, it is unclear what precisely we should do for the 316 factors in our list. In addition, even for the Bayesian approach, the final reject/accept decision still involves the threshold choice. Due to these concerns, we choose not to implement the Bayesian approach. We leave extensions of the basic Bayesian framework that could possibly alleviate the above concerns to future research.

3.7.4 Methods controlling the FDP. Instead of FDR, recent research by Lehmann and Romano (2005) develops methods to directly control the realized FDP. In particular, they propose a step-down method to control for the probability of FDP exceeding a threshold value. Since their definition of type I error (i.e., $P(FDP > \gamma)$, where γ is the threshold FDP value) is different from either FWER or FDR, results based on their methods are not comparable to ours. However, the main conclusion is the same. For instance, when $\gamma = 0.10$ and $\alpha = 0.05$, the benchmark t-statistic is 2.70 (p-value = 0.69%), which is much higher than the conventional cutoff of 1.96. Details are presented in Online Appendix C.

4. Correlation among Test Statistics

Although the BHY method is robust to arbitrary dependence among test statistics, it does not use any information about the dependence structure. Such information, when appropriately incorporated, can be helpful in making the method more accurate (i.e., less stringent). We focus on the type of dependence that can be captured by the Pearson correlation. To generate correlation among test statistics, we focus on the correlation among factor returns. This correlation is likely driven by macroeconomic and market-wide variables. Therefore, in our context, the dependence among test statistics is equivalent to the correlation among factor returns.

Multiple testing corrections in the presence of correlation have been only considered in the recent statistics literature. Existing methods include bootstrap-based permutation tests and direct statistical modeling. Permutation tests resample the entire dataset and construct an empirical distribution for the pool of test statistics.[33] Through resampling, the correlation structure in the data is taken into account and no model is needed. In contrast, direct statistical modeling makes additional distributional assumptions on the data-generating process. These assumptions are usually case dependent as different kinds of correlations are more plausible under different circumstances.[34]

[33] Westfall (1993) and Ge et al. (2003) are the early papers that suggest the permutation resampling approach in multiple testing. Later development of the permutation approach tries to reduce computational burden by proposing efficient alternative approaches. Examples include Lin (2005), Conneely and Boehnke (2007), and Han, Kang, and Eskin (2009).

[34] See Sun and Cai (2009) and Wei et al. (2009).

In addition, recent research in finance explores bootstrap procedures to assess the statistical significance of individual tests.[35] Many of these studies focus on performance evaluation and test whether fund managers exhibit skill. Our approach focuses on the joint distribution of the test statistics (both FWER and FDR depend on the cross-section of t-statistics) and evaluates the significance of each individual factor.

Unfortunately, we do not always observe the time series of factor returns (when a t-statistic is based on long-short strategy returns) or the time series of slopes in cross-sectional regressions (when a t-statistic is based on the slope coefficients in cross-sectional regressions). Because few researchers post their original data, often all we have is the single t-statistic that summarizes the significance of a factor. We propose a novel approach to overcome this missing data problem. It is in essence a "direct modeling approach" but does not require the full information of the return series based on which the t-statistic is constructed. In addition, our approach is flexible enough to incorporate various kinds of distributional assumptions. We expect it to be a valuable addition to the multiple testing literature, especially when only test statistics are observable.

4.1 A model with correlations

For each factor, suppose researchers construct a corresponding long-short trading strategy and normalize the return standard deviation to be $\sigma = 15\%$ per year, which is close to the annual volatility of the market index.[36] In particular, let the normalized strategy return in period t for the i-th discovered strategy be $X_{i,t}$. Then the t-statistic for testing the significance of this strategy is:

$$T_i = (\sum_{t=1}^{N} X_{i,t}/N)/(\sigma/\sqrt{N}).$$

Assuming joint normality and zero serial correlation for strategy returns, this t-statistic has a normal distribution

$$T_i \sim N(\mu_i/(\sigma/\sqrt{N}), 1),$$

where μ_i denotes the population mean of the strategy. The μ_i's are unobservable, and hypothesis testing under this framework amounts to testing $\mu_i > 0$. We assume that each μ_i is an independent draw from the following mixture distribution:

$$\mu_i \sim p_0 I_{\{\mu=0\}} + (1-p_0)\text{Exp}(\lambda),$$

where $I_{\{\mu=0\}}$ is the distribution that has a point mass at zero, Exp(λ) is the exponential distribution that has a mean parameter λ, and p_0 is the probability of

[35] See Efron (1979) for the original work in the statistics literature. For recent finance applications, see Karolyi and Kho (2004), Kosowski et al. (2006), Kosowski, Naik, and Teo (2007), Fama and French (2010), Cao et al. (2013), and Harvey and Liu (2014c).

[36] Notice that this assumption is not necessary for our approach. Fixing the standard deviations of different strategies eliminates the need to separately model them, which can be done through a joint modeling of the mean and variance of the cross-section of returns. See Harvey and Liu (2014a) for further discussions on this.

drawing from the point mass distribution. This mixture distribution assumption is the core component for Bayesian multiple testing and succinctly captures the idea of hypothesis testing in the traditional frequentist's view: while there is a range of possible values for the means of truly profitable strategies, a proportion of strategies should have a mean that is indistinguishable from zero. The exponential assumption is not essential for our model as more sophisticated distributions (e.g., a Gamma distribution featuring two free parameters) can be used. We use the exponential distribution for its simplicity[37] and, perhaps more importantly, because it is consistent with the intuition that more profitable strategies are less likely to exist. An exponential distribution captures this feature by having a monotonically decreasing probability density function.

Next, we incorporate correlations into the above framework. Among the various sources of correlations, the cross-sectional correlations among contemporaneous returns are the most important for us to take into account. These correlations are likely induced by a response to common macroeconomic or market shocks. Other kinds of correlations can be easily embedded into our framework as well.[38]

As a starting point, we assume that the contemporaneous correlation between two strategies' returns is ρ. The noncontemporaneous correlations are assumed to be zero. That is,

$$\mathrm{Corr}(X_{i,t}, X_{j,t}) = \rho, \quad i \neq j,$$

$$\mathrm{Corr}(X_{i,t}, X_{j,s}) = 0, \quad t \neq s.$$

Finally, to incorporate the impact of hidden tests, we assume that M factors are tried, but only factors that exceed a certain t-statistic threshold are published. We set the threshold t-statistic at 1.96 and focus on the subsample of factors that have a t-statistic larger than 1.96. However, as shown in Appendix A, factors with marginal t-statistics (i.e., t-statistics just above 1.96) are less likely to be published than those with larger t-statistics. Therefore, our subsample of published t-statistics only covers a fraction of t-statistics above 1.96 for tried factors. To overcome this missing data problem, we assume that our sample covers a fraction r of t-statistics in between 1.96 and 2.57 and that all t-statistics above 2.57 are covered. We augment the existing t-statistic sample to construct the full sample. For instance, when $r = 1/2$, we simply duplicate the sample of t-statistics in between 1.96 and 2.57 and maintain the sample of t-statistics above 2.57 to construct the full

[37] As shown later, we need to estimate the parameters in the mixture model based on our t-statistics sample. An overparameterized distribution for the continuous distribution in the mixture model, albeit flexible, may result in imprecise estimates. We therefore use the simple one-parameter exponential distribution family.

[38] To incorporate the serial correlation for individual strategies, we can model them as simple autoregressive processes. See Harvey and Liu (2014a) for further discussion of the kinds of correlation structures that our model is able to incorporate. See Sun and Cai (2009) for an example that models the spatial dependence among the null hypotheses.

sample. For the baseline case, we set $r = 1/2$, consistent with the analysis in Appendix A. We try alternative values of r to determine how the results change.[39]

Given the correlation structure and the sampling distribution for the means of returns, we can fully characterize the distributional properties of the cross-section of returns. We can also determine the distribution for the cross-section of t-statistics as they are functions of returns. Based on our sample of t-statistics for published research, we match key sample statistics with their population counterparts in the model.

The sample statistics we choose to match are the quantiles of the sample of t-statistics and the sample size (i.e., the total number of discoveries). Two concerns motivate us to use quantiles. First, sample quantiles are less susceptible to outliers compared to means and other moment-related sample statistics. Our t-statistic sample does have a few influential observations, and we expect quantiles to be more useful descriptive statistics than the mean and the standard deviation. Second, simulation studies show that quantiles in our model are more sensitive to changes in parameters than other statistics. To offer a more efficient estimation of the model, we choose to focus on quantiles.

In particular, the quantities we choose to match and their values for the baseline sample (i.e., $r = 1/2$) are given by:

$$\begin{cases} \widehat{T} = \text{Total number of discoveries} = 353, \\ \widehat{Q}_1 = \text{The 20th percentile of the sample of } t\text{-statistics} = 2.39, \\ \widehat{Q}_2 = \text{The 50th percentile of the sample of } t\text{-statistics} = 3.16, \\ \widehat{Q}_3 = \text{The 90th percentile of the sample of } t\text{-statistics} = 6.34. \end{cases}$$

These three quantiles are representative of the spectrum of quantiles and can be shown to be most sensitive to parameter changes in our model. Fixing the model parameters, we can also obtain the model implied sample statistics T, Q_1, Q_2, and Q_3 through simulations.[40] The estimation works by seeking to find the set of parameters that minimizes the following objective function:

$$D(\lambda, p_0, M, \rho) = w_0(T - \widehat{T})^2 + \sum_{i=1}^{3} w_i(Q_i - \widehat{Q}_i)^2,$$

where w_0 and $\{w_i\}_{i=1}^{3}$ are the weights associated with the squared distances. Motivated by the optimal weighting for the generalized method of moments (GMM) estimators, we set these weights at $w_0 = 1$ and $w_1 = w_2 = w_3 = 10,000$.

---

39 Our choice of the threshold t-statistic is smaller than the 2.57 threshold in Appendix A. This allows us to observe false discoveries that overcome the threshold more frequently than under 2.57. This is important for the estimation of p_0 in the model. For more details on the selection of the threshold t-statistic, see Harvey and Liu (2014a).

40 Model implied quantiles are difficult (and most likely infeasible) to calculate analytically. We obtain them through simulations. In particular, for a fixed set of parameters, we simulate 5,000 independent samples of t-statistics. For each sample, we calculate the four summary statistics. The median of these summary statistics across the 5,000 simulations is taken as the model implied statistics.

They can be shown to have the same magnitude as the inverses of the variances of the corresponding model implied sample statistics across a wide range of parameter values and should help improve estimation efficiency.[41]

We estimate the three parameters (λ, p_0, and M) in the model and choose to calibrate the correlation coefficient ρ. In particular, for a given level of correlation ρ, we numerically search for the model parameters (λ, p_0, M) that minimize the objective function $D(\lambda, p_0, M, \rho)$.

We choose to calibrate the amount of correlation because the correlation coefficient is likely to be weakly identified in this framework. Ideally, to have a better identification of ρ, we would like to have t-statistics that are generated from samples that have varying degrees of overlap.[42] We do not allow heterogeneity in sample periods in either our estimation framework (i.e., all t-statistics are generated from samples that cover the same period) or our data (we do not record the specific period for which the t-statistic is generated). As a result, our results are best interpreted as the estimated t-statistic thresholds for a hypothetical level of correlation.

To investigate how correlation affects multiple testing, we follow an intuitive simulation procedure. In particular, fixing λ, p_0, and M at their estimates, we know the data-generating process for the cross-section of returns. Through simulations, we are able to calculate the previously defined type I error rates (i.e., FWER and FDR) for any given threshold t-statistic. We search for the optimal threshold t-statistic that exactly achieves a prespecified error rate.

4.2 Results

Our estimation framework assumes a balanced panel with M factors and N periods of returns. We need to assign a value to N. Published papers usually cover a period ranging from twenty to fifty years. In our framework, the choice of N does not affect the distribution of T_i under the null hypothesis (i.e., $\mu_i = 0$) but will affect T_i under the alternative hypothesis (i.e., $\mu_i > 0$). When μ_i is different from zero, T_i has a mean of $\mu_i/(\sigma/\sqrt{N})$. A larger N reduces the noise in returns and makes it more likely for T_i to be significant. To be conservative (i.e., less likely to generate significant t-statistics under the alternative hypotheses), we set N at 240 (i.e., twenty years). Other specifications of N change the estimate of λ but leave the other parameters almost intact. In particular, the threshold t-statistics are little changed for alternative values of N.

The results are presented in Table 5. Across different correlation levels, λ (the mean parameter for the exponential distribution that represents the mean

---

41 We do not pursue a likelihood-based estimation. Given that we have more than a thousand factors and each of them is associated with an indicator variable that is missing, the likelihood function involves high-dimensional integrals and is difficult to optimize. This leads us to a GMM-based approach.

42 Intuitively, t-statistics that are based on similar sample periods are more correlated than t-statistics that are based on distinct sample periods. Therefore, the degree of overlap in sample period helps identify the correlation coefficient. See Ferson and Chen (2013) for a similar argument on measuring the correlations among fund returns.

Table 5
Estimation results: A model with correlations

Panel A: r = 1/2 (baseline)

				t-statistic			
ρ	p_0	λ(%)	M	FWER(5%)	FWER(1%)	FDR(5%)	FDR(1%)
0	0.396	0.550	1,297	3.89	4.28	2.16	2.88
0.2	0.444	0.555	1,378	3.91	4.30	2.27	2.95
0.4	0.485	0.554	1,477	3.81	4.23	2.34	3.05
0.6	0.601	0.555	1,775	3.67	4.15	2.43	3.09
0.8	0.840	0.560	3,110	3.35	3.89	2.59	3.25

Panel B: r = 2/3 (more unobserved tests)

0	0.683	0.550	2,458	4.17	4.55	2.69	3.30
0.2	0.722	0.551	2,696	4.15	4.54	2.76	3.38
0.4	0.773	0.552	3,031	4.06	4.45	2.80	3.40
0.6	0.885	0.562	4,339	3.86	4.29	2.91	3.55
0.8	0.922	0.532	5,392	3.44	4.00	2.75	3.39

We estimate the model with correlations. r is the assumed proportion of missing factors with a t-statistic between 1.96 and 2.57. Panel A shows the results for the baseline case in which $r = 1/2$, and panel B shows the results for the case in which $r = 2/3$. ρ is the correlation coefficient between two strategy returns in the same period. p_0 is the probability of having a strategy that has a mean of zero. λ is the mean parameter of the exponential distribution for the monthly means of the true factors. M is the total number of trials.

returns for true factors) is consistently estimated at 0.55% per month. This corresponds to an annual factor return of 6.6%. Therefore, we estimate the average mean returns for truly significant factors to be 6.6% per annum. Given that we standardize factor returns by an annual volatility of 15%, the average annual Sharpe ratio for these factors is 0.44 (or monthly Sharpe ratio of 0.13).[43]

For the other parameter estimates, both p_0 and M are increasing in ρ. Focusing on the baseline case in panel A and at $\rho = 0$, we estimate that researchers have tried $M = 1{,}297$ factors and 60.4% $(= 1 - 0.396)$ are true discoveries. When ρ is increased to 0.60, we estimate that a total of $M = 1{,}775$ factors have been tried and around 39.9% $(= 1 - 0.601)$ are true factors.

Turning to the estimates of threshold t-statistics and focusing on FWER, we see that they are not monotonic in the level of correlation. Intuitively, two forces are at work in driving these threshold t-statistics. On the one hand, both p_0 and M are increasing in the level of correlation. Therefore, more factors—both in absolute value and in proportion—are drawn from the null hypothesis. To control the occurrences of false discoveries based on these factors, we need a higher threshold t-statistic. On the other hand, a higher correlation among test statistics reduces the required threshold t-statistic. In the extreme case when all test statistics are perfectly correlated, we do not need multiple testing adjustment at all. These two forces work against each other and result in the nonmonotonic pattern for the threshold t-statistics under FWER. For FDR,

43 Our estimates are robust to the sample percentiles that we choose to match. For instance, fixing the level of correlation at 0.2, when we use the 10th together with the 50th and 90th percentiles of the sample of t-statistics, our parameter estimate is $(p_0, \lambda, M) = (0.390, 0.548, 1{,}287)$. Alternatively, when we use the 80th together with the 20th and 50th percentiles of the sample of t-statistics, our parameter estimate is $(p_0, \lambda, M) = (0.514, 0.579, 1{,}493)$. Both estimates are in the neighborhood of our baseline model estimates.

it appears that the impact of larger p_0 and M dominates so that the threshold t-statistics are increasing in the level of correlation.

Across various correlation specifications, our estimates show that in general a t-statistic of 3.9 and 3.0 is needed to control FWER at 5% and FDR at 1%, respectively.[44] Notice that these numbers are not far away from our previous estimates of 3.78 (Holm adjustment that controls FWER at 5%) and 3.38 (BHY adjustment that controls FDR at 1%). However, these similar numbers are generated through different mechanisms. Our current estimate assumes a certain level of correlation among returns and relies on an estimate of more than 1,300 for the total number of factor tests. On the other hand, our previous calculation assumes that the 316 published factors are all the factors that have been tried but does not specify a correlation structure.

4.3 How large is ρ?

Our sample has limitations in making a direct inference on the level of correlation. To give some guidance, we provide indirect evidence on the plausible levels of ρ.

First, the value of the optimized objective function sheds light on the level of ρ. Intuitively, a value of ρ that is more consistent with the data-generating process should result in a lower optimized objective function. Across the various specifications of ρ in Table 5, we find that the optimized objective function reaches its lowest point when $\rho = 0.2$. Therefore, our t-statistic sample suggests a low level of correlation. However, this evidence is only suggestive given the weak identification of ρ in our model.

Second, we draw on external data source to provide inference. In particular, we analyze the S&P CAPITAL IQ database, which includes detailed information on the time-series of returns of over 400 factors for the U.S. equity market. We estimate the average pairwise correlation among these factors to be 0.15 for the 1985–2014 period.

Finally, existing studies in the literature provide guidance on the level of correlation. McLean and Pontiff (2015) estimate the correlation among anomaly returns to be around 0.05. Green, Hand, and Zhang (2013a) focus on accounting-based factors and find the average correlation to be between 0.06 and 0.20. Focusing on mutual fund returns, Barras, Scaillet, and Wermers (2010) argue for a correlation of zero among fund returns (i.e., excess returns against benchmark factors), while Ferson and Chen (2013) calibrate this number to be between 0.04 and 0.09.

Overall, we believe that the average correlation among factor returns is in the neighborhood of 0.20.

---

[44] To save space, we choose not to discuss the performance of our estimation method. Harvey and Liu (2014a) provide a detailed simulation study of our model.

4.4 How many true factors are there?

The number of true discoveries using our method seems high given that most of us believe as a priori that there are only a handful of true systematic risk factors. However, many of these factors that our method deems statistically true have tiny Sharpe ratios. For example, around 70% of them have a Sharpe ratio that is less than 0.5 per annum. From a modeling perspective, we impose a monotonic exponential density for the mean returns of true factors. Hence, by assumption, the number of discoveries will be decreasing in the mean return.

Overall, statistical evidence can only get us so far in reducing the number of false discoveries. This is a limitation not only to our framework but also probably in any statistical framework that relies on individual p-values. To see this, suppose the smallest t-statistic among true risk factors is 3.0 and assume our sample covers fifty risk factors that all have a t-statistic above 3.0. Then based on statistical evidence only, it is impossible to rule out any of these fifty factors from the list of true risk factors.

We agree that a further scrutiny of the factor universe is a valuable exercise. There are at least two routes we can take. One route is to introduce additional testable assumptions that a systematic risk factor has to satisfy to claim significance. Pukthuanthong and Roll (2014) use the principle components of the cross-section of realized returns to impose such assumptions. The other route is to incrementally increase the factor list by allowing different factors to crowd each other out. Harvey and Liu (2014c) provide such a framework. We expect both lines of research to help in culling the number of factors.

5. Conclusion

At least 316 factors have been tested to explain the cross-section of expected returns. Most of these factors have been proposed over the last ten years. Indeed, Cochrane (2011) refers to this as “a zoo of new factors.” Our paper argues that it is a serious mistake to use the usual statistical significance cutoffs (e.g., a t-statistic exceeding 2.0) in asset pricing tests. Given the plethora of factors, and the inevitable data mining, many of the historically discovered factors would be deemed “significant” by chance.

There is an important philosophical issue embedded in our approach. Our threshold cutoffs increase through time as more factors are data mined. However, data mining is not new. Why should we have a higher threshold for today’s data mining than for data mining in the 1980s? We believe there are three reasons for tougher criteria today. First, the low-hanging fruit has already been picked. That is, the rate of discovering a true factor has likely decreased. Second, there is a limited amount of data. Indeed, there is only so much you can do with the CRSP database. In contrast, in particle physics, it is routine to create trillions of new observations in an experiment. We do not have that luxury in finance. Third, the cost of data mining has dramatically decreased. In the past, data collection and estimation were time intensive, so it was more likely that only factors with the highest priors—potentially based on economic first principles—were tried.

Our paper presents three conventional multiple testing frameworks and proposes a new one that particularly suits research in financial economics. While these frameworks differ in their assumptions, they are consistent in their conclusions. We argue that a newly discovered factor today should have a t-statistic that exceeds 3.0. We provide a time-series of recommended "cutoffs" from the first empirical test in 1967 through to present day. Many published factors fail to exceed our recommended cutoffs.

While a t-statistic of 3.0 (which corresponds to a p-value of 0.27%) seems like a very high hurdle, we also argue that there are good reasons to expect that 3.0 is too low. First, we only count factors that are published in prominent journals and we sample only a small fraction of the working papers. Second, there are surely many factors that were tried by empiricists, failed, and never made it to publication or even a working paper. Indeed, the culture in financial economics is to focus on the discovery of new factors. In contrast with other fields, such as medical science, it is rare to publish replication studies focusing on only existing factors. Given that our count of 316 tested factors is surely too low, this means the t-statistic cutoff is likely even higher.[45]

Should a t-statistic of 3.0 be used for every factor proposed in the future? Probably not. A case can be made that a factor developed from first principles should have a lower threshold t-statistic than a factor that is discovered as a purely empirical exercise. Nevertheless, a t-statistic of 2.0 is no longer appropriate—even for factors that are derived from theory.

In medical research, the recognition of the multiple testing problem has led to the disturbing conclusion that "most claimed research findings are false" (Ioannidis (2005)). Our analysis of factor discoveries leads to the same conclusion – many of the factors discovered in the field of finance are likely false discoveries: of the 296 published significant factors, 158 would be considered false discoveries under Bonferonni, 142 under Holm, 132 under BHY (1%), and 80 under BHY (5%). In addition, the idea that there are so many factors is inconsistent with the principal component analysis, where, perhaps there are five "statistical" common factors driving time-series variation in equity returns (Ahn, Horenstein, and Wang 2012).

The assumption that researchers follow the rules of classical statistics (e.g., randomization, unbiased reporting) is at odds with the notion of individual incentives, ironically, one of the fundamental premises in economics. Importantly, the optimal amount of data mining is not zero since some data mining produces knowledge. The key, as argued by Glaeser (2008), is to design appropriate statistical methods to adjust for biases, not to eliminate research

---

45 In astronomy and physics, even higher threshold t-statistics are often used to control for testing multiplicity. For instance, the high profile discovery of Higgs Boson has a t-statistic of more than 5 (p-value less than 0.0001%). See ATLAS Collaboration (2012), CMS Collaboration (2012), and Harvey and Liu (2014b).

Table 6
Factor list: Factors sorted by year

Reference	Factor	#	Reference	Factor	#
Sharpe (1964)	market return	T	Constantinides (1982)	individual consumer's wealth	T
Lintner (1965)	market return	T	Basu (1983)	EP ratio	C8
Mossin (1966)	market return	T	Adler and Dumas (1983)	FX rate change	T
Douglas (1967)	total volatility	C1	Arbel, Carvell, and Strebel (1983)	institutional holding‡	
Heckerman (1972)	market return	T	Hawkins, Chamberlin, and Daniel (1984)	earnings expectations‡	
	relative prices of cons. goods	T	McConnell and Sanger (1984)	new listings announcement‡	
[1]Black, Jensen, and Scholes (1972)	market return		Chan, Chen, and Hsieh (1985)	market return†	
Black (1972)	market return	T		industrial production growth	F5
Merton (1973)	state variables investment opps.	T		change in expected inflation*	F6
Fama and MacBeth (1973)	market return	F1		unanticipated inflation	F7
	beta squared*	F2		credit premium	F8
	idiosyncratic volatility*	C2		term structure*	F9
Rubinstein (1973)	high-order market return	T	De Bondt and Thaler (1985)	long-term return reversal	C9
Solnik (1974)	world market return	T	Cox, Ingersoll, and Ross (1985)	Δ investment opportunities	T
Rubinstein (1974)	individual investor resources	T	Amihud and Mendelson (1986)	transaction costs	T
Gupta and Ofer (1975)	earnings growth expectations	C3	Constantinides (1986)	transaction costs	T
Kraus and Litzenberger (1976)	market return†		Stulz (1986)	expected inflation	T
	squared market return*	F3	Sweeney and Warga (1986)	long-term interest rate	F10
Basu (1977)	PE ratio	C4	Chen, Roll, and Ross (1986)	industrial production growth†	
Lucas (1978)	marginal rate of substitution	T		credit premium†	
Litzenberger and Ramaswamy (1979)	dividend yield	C5		term structure†	
	market return†			unanticipated inflation†	
Breeden (1979)	real consumption growth	T		change in oil prices*	F11
Jarrow (1980)	short-sale restrictions	T	Bhandari (1988)	debt-to-equity ratio	C10
[2]Fogler, John, and Tipton (1981)	market return†‡		Bauman and Dowen (1988)	long-term growth forecasts‡	
	Treasury bond return‡		Breeden, Gibbons, and Litzenberger (1989)	consumption growth	F12
	corporate bond return‡		Amihud and Mendelson (1989)	illiquidity	C11
Oldfield and Rogalski (1981)	Treasury-bill return	F4	Ou and Penman (1989)	predicted earnings change	C12
Stulz (1981)	world consumption	T	Jegadeesh (1990)	return predictability	C13
Mayshar (1981)	transaction costs	T			
Banz (1981)	firm size	C6			
Figlewski (1981)	short interest	C7			

(continued)

Table 6
Continued

Reference	Factor	#	Reference	Factor	#
Ferson and Harvey (1991)	market return†		Elton, Gruber, and Blake (1995)	change in expected inflation	F22
	consumption growth†			change in expected GNP	F23
	credit spread†		Spiess and Affleck-Graves (1999)	seasoned equity offerings‡	
	Δ slope of the yield curve	F13	Chan, Foresi, and Lang (1996)	money growth	F24
	unexpected inflation†		Cochrane (1996)	returns on physical inv.	F25
	real short rate	F14	Campbell (1996)	market return†	
[3]Fama and French (1992)	size	F15		labor income	F26
	value	F16		dividend yield†	
Chopra, Lakonishok, and Ritter (1992)	return momentum‡			interest rate†	
Holthausen and Larcker (1992)	predicted return signs‡			term structure†	
Jegadeesh and Titman (1993)	return momentum	C14	Jagannathan and Wang (1996)	market return†	
Elton et al. (1993)	returns on S&P stocks‡			slope of yield curve†	
	returns on non-S&P stocks‡			labor income†	
[4]Bansal and Viswanathan (1993)	high-order equity & bond returns‡		La Porta (1996)	earnings forecasts	C18
Fama and French (1993)	market return†		Lev and Sougiannis (1996)	R&D capital	C19
	size†		Sloan (1996)	accruals	C20
	value†		Womack (1996)	buy recommendations	C21
	term structure†			sell recommendations	C22
	credit risk†		Erb, Harvey, and Viskanta (1996)	credit rating	C23
[5]Ferson and Harvey (1993)	world equity return‡		Brennan and Subrahmanyam (1996)	illiquidity	C24
	change in weighted exchange rate‡		[6]Chapman (1997)	nonlinear fn. of cons. growth‡	
	Δ LT inflation expectations‡		[7]Fung and Hsieh (1997)	opportunistic style return‡	
	weighted real short-term rate‡			global/macro style return‡	
	change in oil price†‡			value style return‡	
	change in TED spread‡			trend following style return‡	
	Δ in G-7 industrial production‡			distressed inv. style return‡	
	unexpected G-7 inflation‡		Carhart (1997)	size†	
Ferson and Harvey (1994)	world equity return	F17		value†	
	change in weighted FX rate*	F18		market return†	
	Δ LT inflation expectations*	F19		momentum	F27
	change in oil price*†		Botosan (1997)	disclosure level	C25
Bossaerts and Dammon (1994)	tax rate for capital gains	F20	Ackert and Athanassakos (1997)	earnings forecast uncertainty	C26
	tax rate for dividend	F21	Daniel and Titman (1997)	size†	
Loughran and Ritter (1995)	new public stock issuance	C15		value†	
Michaely, Thaler, and Womack (1995)	dividend initiations	C16			
	dividend omissions	C17			

(*continued*)

Table 6
Continued

Reference	Factor	#	Reference	Factor	#
Beneish (1997)	earnings management likelihood‡		Griffin and Lemmon (2002)	distress risk	C45
Loughran and Vijh (1997)	corporate acquisitions	C27	Diether, Malloy, and Scherbina (2002)	analyst dispersion	C46
Brennan, Chordia, and Subrahmanyam (1998)	size†		Chen, Hong, and Stein (2002)	breadth of ownership	C47
	book-to-market ratio†		Easley, Hvidkjaer, and O'Hara (2002)	information risk	C48
	momentum†		Jones and Lamont (2002)	short-sale constraints	C49
	trading volume	C28	Penman and Zhang (2002)	earnings sustainability	C50
Abarbanell and Bushee (1998)	fundamental analysis‡		Amihud (2002)	market illiquidity	F33
Frankel and Lee (1998)	firm fundamental value‡		Vassalou (2003)	GDP growth news	F34
Dichev (1998)	bankruptcy risk	C29	Pastor and Stambaugh (2003)	market liquidity	F35
Datar, Naik, and Radcliffe (1998)	illiquidity	C30	Ali, Hwang, and Trombley (2003)	idiosyncratic return volatility†	
Ferson and Harvey (1999)	expected portfolio return	F28		transaction costs†	
Moskowitz and Grinblatt (1999)	industry momentum	C31		investor sophistication†	
Spiess and Affleck-Graves (1999)	debt offerings‡		Gompers, Ishii, and Metrick (2003)	shareholder rights	C51
Heaton and Lucas (2000)	entrepreneur income	F29	Doyle, Lundholm, and Soliman (2003)	excluded expenses	C52
Harvey and Siddique (2000)	coskewness	F30	Fairfield, Whisenant, and Yohn (2003)	growth in LT net operating assets	C53
Lee and Swaminathan (2000)	trading volume	C32	Rajgopal, Shevlin, and Venkatachalam (2003)	order backlog	C54
Asness, Porter, and Stevens (2000)	intra-industry size	C33	Watkins (2003)	return consistency	C55
	intra-industry value	C34	Jacobs and Wang (2004)	idiosyncratic consumption	F36
	intra-industry CF/P	C35	Campbell and Vuolteenaho (2004)	cash-flow news	F37
	intra-industry Δ% # employees	C36		discount rate news	F38
	intra-industry momentum	C37	[8]Vanden (2004)	market return†	
Piotroski (2000)	financial statement infor.	C38		index option returns	F39
Lettau and Ludvigson (2001)	consumption growth†		Vassalou and Xing (2004)	default risk	F40
	consumption-wealth ratio	F31	Brennan, Wang, and Xia (2004)	real interest rate	F41
Chordia, Subrahmanyam, and Anshuman (2001)	level of liquidity	C39		maximum Sharpe ratio portfolio	F42
	variability of liquidity	C40	Teo and Woo (2004)	return reversals at the style level	F43
Lamont, Polk, and Saa-Requejo (2001)	financial constraints	C41	Eberhart, Maxwell, and Siddique (2004)	unexpected change in R&D	C56
Fung and Hsieh (2001)	straddle return‡		George and Hwang (2004)	52-week high	C57
Barber et al. (2001)	consensus recommendations*		Jegadeesh et al. (2004)	analysts' recommendations	C58
Dichev and Piotroski (2001)	bond rating changes	C42	Ofek, Richardson, and Whitelaw (2004)	put-call parity	C59
Elgers, Lo, and Pfeiffer (2001)	analysts' forecasts	C43	Titman, Wei, and Xie (2004)	abnormal capital investment	C60
Gompers and Metrick (2001)	institutional ownership	C44	Hirshleifer et al. (2004)	balance sheet optimism	C61
Dittmar (2002)	market return†		Parker and Julliard (2005)	LT consumption growth	F44
	squared market return†		Bansal, Dittmar, and Lundblad (2005)	long-run consumption	F45
	labor income growth†				
	squared labor income growth	F32			

(*continued*)

Table 6
Continued

Reference	Factor	#	Reference	Factor	#
Lustig and Van Nieuwerburgh (2005)	housing price ratio	F46	Brammer, Brooks, and Pavelin (2006)	environment indicator*	C78
Cremers and Nair (2005)	external corporate governance	C62		employment indicator*	C79
	internal corporate governance	C63		community indicator*	C80
[9]Acharya and Pedersen (2005)	market return†		Daniel and Titman (2006)	intangible information	C81
	market liquidity*	F47	Fama and French (2006)	profitability	C82
	individual stock liquidity	C64		investment*	C83
Hou and Moskowitz (2005)	price delay	C65		book-to-market†	
Anderson, Ghysels, and Juergens (2005)	heterogeneous beliefs	C66	Bradshaw, Richardson, and Sloan (2006)	net financing	C84
Nagel (2005)	short-sale constraints	C67	Cen, Wei, and Zhang (2006)	forecasted earnings per share	C85
Asquith, Pathak, and Ritter (2005)	short-sale constraints	C68	Franzoni and Marin (2006)	pension plan funding	C86
Gu (2005)	patent citation	C69	Gettleman and Marks (2006)	acceleration	C87
Jiang, Lee and Zhang (2005)	information uncertainty	C70	Narayanamoorthy (2006)	unexpected earnings' autocorr.	C88
Lev, Nissim, and Thomas (2005)	adjusted R&D	C71	Boudoukh et al. (2007)	payout yield	F62
Lev, Sarath, and Sougiannis (2005)	R&D reporting biases	C72	Balvers and Huang (2007)	productivity	F63
Mohanram (2005)	growth index	C73		capital stock	F64
[10]Vanden (2006)	market return†		Jagannathan and Wang (2007)	4th Q to 4th Q cons. growth	F65
	index option return†		Avramov et al. (2007)	credit rating	C89
	market × option return‡		Shu (2007)	trader composition	C90
Gomes, Yaron, and Zhang (2006)	financing frictions	F48	Baik and Ahn (2007)	change in order backlog	C91
Li, Vassalou, and Xing (2006)	inv. growth (IG) households*	F49	Brown and Rowe (2007)	firm productivity	C92
	IG nonfinancial corporates	F50	Doran, Fodor, and Peterson (2007)	insider forecasts of firm vol.	C93
	IG noncorporate business	F51	Head, Smith, and Wilson (2007)	ticker symbol	C94
	IG financial firms	F52	Gourio (2007)	earnings cyclicality	F66
[11]Chung, Johnson, and Schill (2006)	3rd-10th power market return‡		Kumar et al. (2008)	market volatility innovation	F67
Whited and Wu (2006)	financial constraints	C74		firm age	C95
Ang et al. (2006)	downside risk	F53		market return†	
Ang et al. (2006)	systematic volatility	F54		market vol. × firm age	C96
	idiosyncratic volatility	C75	Adrian and Rosenberg (2008)	short-run market volatility	F68
Baker and Wurgler (2006)	investor sentiment	F55		long-run market volatility	F69
Kumar and Lee (2006)	retail investor sentiment	F56	Xing (2008)	investment growth	F70
Yogo (2006)	durable & nondur. cons. growth	F57	Korniotis (2008)	mean consumption growth	F71
Lo and Wang (2006)	market return†			variance of consumption growth*	F72
	trading volume	F58		mean habit growth	F73
Sadka (2006)	liquidity	F59		variance of habit growth	F74
Chordia and Shivakumar (2006)	earnings	F60	Korajczyk and Sadka (2008)	liquidity	F75
Liu (2006)	liquidity	F61	Guo and Savickas (2008)	country-level idiosyncratic vol.	C97
Anderson and Garcia-Feijóo (2006)	capital investment	C76	Campbell, Hilscher, and Szilagyi (2008)	distress	C98
Hou and Robinson (2006)	industry concentration	C77	Garlappi, Shu, and Yan (2008)	shareholder advantage	C99
				implied market value from KMV	C100

(continued)

Table 6
Continued

Reference	Factor	#	Reference	Factor	#
Cooper, Gulen, and Schill (2008)	asset growth	C101	Barber, Odean, and Zhu (2009)	order imbalance	C125
Pontiff and Woodgate (2008)	share issuance	C102	Cremers, Halling, and Weinbaum (2010)	market volatility and jumps	F85
Brandt et al. (2008)	earnings announcement return‡		Hirshleifer and Jiang (2010)	market mispricing	F86
Cohen and Frazzini (2008)	firm economic links	C103	Boyer, Mitton, and Vorkink (2010)	idiosyncratic skewness	C126
Fabozzi, Ma, and Oliphant (2008)	sin stock	C104	Cooper, Gulen, and Ovtchinnikov (2010)	political contributions	C127
Gu and Lev (2011)	goodwill impairment	C105	Tuzel (2010)	real estate holdings	C128
Gu, Wang, and Ye (2008)	information in order backlog	C106	Amaya et al. (2011)	realized skewness	C129
Lehavy and Sloan (2008)	investor recognition	C107		realized kurtosis	C130
Soliman (2008)	DuPont analysis	C108	An, Bhojraj, and Ng (2010)	excess multiple	C131
Hvidkjaer (2008)	small trades	C109	Armstrong, Banerjee, and Corona (2010)	firm information quality	C132
Brennan and Li (2008)	idiosyncratic S&P 500 return	F76	Cao and Xu (2010)	long-run idiosyncratic vol.	C133
Da (2009)	cash flow cova. with cons.	F77	Easley, Hvidkjaer, and O'Hara (2010)	private information	F87
	cash flow duration	F78	Hameed, Huang, and Mian (2010)	intra-industry return reversals	C134
Livdan, Sapriza, and Zhang (2009)	financial constraints		Menzly and Ozbas (2010)	related industry returns	C135
Malloy, Moskowitz, and Vissing-Jorgensen (2009)	LT stockholder cons. growth	F79	Papanastasopoulos, Thomakos, and Wang (2010)	earnings to equity holders	C136
Cremers, Nair, and John (2009)	takeover likelihood	F80		net cash to equity holders	C137
Chordia, Huh, and Subrahmanyam (2009)	illiquidity	F81	Simutin (2010)	excess cash	C138
Da and Warachka (2009)	cash flow	F82	Huang et al. (2012)	extreme downside risk	C139
Ozoguz (2009)	investors' beliefs*	F83	Xing, Zhang, and Zhao (2010)	volatility smirk	C140
	investors' uncertainty	F84	George and Hwang (2010)	exposure financial distress costs	
Fang and Peress (2009)	media coverage	C110	Berkman, Jacobsen, and Lee (2011)	rare disasters	F88
Avramov et al. (2009)	financial distress	C111	[12]Kapadia (2011)	distress risk‡	
Fu (2009)	idiosyncratic volatility	C112	Hou, Karolyi, and Kho (2011)	momentum†	
Hahn and Lee (2009)	debt capacity	C113		cash flow-to-price	F89
Bali and Hovakimian (2009)	realized-implied vol. spread	C114	Li (2011)	R&D investment	C141
	call-put implied vol. spread	C115		financial constraints†	
Chandrashekar and Rao (2009)	productivity of cash	C116	Bali, Cakici, and Whitelaw (2011)	extreme stock returns	C142
Chemmanur and Yan (2009)	advertising	C117	Yan (2011)	jumps individual stock returns	C143
Da and Warachka (2009)	analyst forecasts optimism	C118	Edmans (2011)	intangibles	C144
Gokcen (2009)	information revelation	C119	[13]Chen, Novy-Marx, and Zhang (2011)	market return†	
Gow and Taylor (2009)	earnings volatility	C120		investment portfolio return	F90
Huang (2009)	cash-flow volatility	C121		ROE portfolio return	F91
Korniotis and Kumar (2009)	local unemployment	C122	Akbas, Armstrong, and Petkova (2011)	volatility of liquidity	C145
	local housing collateral	C123	Jiang and Sun (2011)	dispersion in beliefs	C146
Nguyen and Swanson (2009)	efficiency score	C124	Han and Zhou (2011)	credit default swap spreads	C147

(continued)

Table 6
Continued

Reference	Factor	#	Reference	Factor	#
Eisfeldt and Papanikolaou (2011)	organizational capital	C148	Lioui and Maio (2012)	future growth opp. cost of money	F106
Balachandran and Mohanram (2011)	residual income	C149	Gârleanu, Kogan, and Panageas (2012)	inter-cohort cons. differences	T
Bandyopadhyay, Huang, and Wirjanto (2010)	accrual volatility	C150	Hu, Pan, and Wang (2012)	market-wide liquidity	F107
Callen and Lyle (2011)	implied cost of capital	C151	Conrad, Dittmar, and Ghysels (2013)	stock skewness	C166
Callen, Khan, and Lu (2013)	nonaccounting infor. quality	C152	Baltussen, Van Bekkum, and Van der Grient (2012)	expected return uncertainty	C167
	accounting infor. quality	C153	Zhao (2012)	information intensity	C168
Chen, Kacperczyk, and Ortiz-Molina (2011)	labor unions	C154	Friewald, Wagner, and Zechner (2012)	credit risk premia	C169
Da, Liu, and Schaumburg (2011)	overreaction to nonfundamentals	C155	Garcia and Norli (2012)	geographic dispersion	C170
Drake, Rees, and Swanson (2011)	short interest	C156	Kim, Pantzalis, and Park (2012)	political geography	C171
Hafzalla, Lundholm, and Van Winkle (2011)	percent total accrual	C157	Johnson and So (2012)	option to stock volume ratio	C172
Hess, Kreutzmann, and Pucker (2011)	projected earnings accuracy‡		Palazzo (2012)	cash holdings	C173
Imrohoroglu and Tuzel (2011)	firm productivity	C158	Donangelo (2012)	labor mobility	C174
Landsman et al. (2011)	really dirty surplus	C159	Wang (2012)	debt covenant protection	C175
Li (2011)	earnings forecast	C160	Chen and Strebulaev (2012)	stock cash-flow sensitivity	C176
Nyberg and Pöyry (2011)	asset growth	C161	Li (2012)	jump beta	F108
Ortiz-Molina and Phillips (2011)	real asset liquidity	C162	[15]Ferson, Nallareddy, and Xie (2012)	long-run cons. growth‡	
Patatoukas (2011)	customer-base concentration	C163		short-run cons. growth‡	
Thomas and Zhang (2011)	tax expense surprises	C164		cons. growth volatility‡	
Wahlen and Wieland (2011)	predicted earnings increase score‡		Ang, Bali, and Cakici (2012)	change in call implied vol.	C177
Garlappi and Yan (2011)	shareholder recovery			change in put implied vol.	C178
Savov (2011)	garbage growth	F92	Bazdresch, Belo and Lin (2012)	firm hiring rate	C179
Adrian, Etula, and Muir (2012)	financial intermediary's wealth	F93	Cohen and Lou (2012)	infor. processing complexity	C180
Campbell et al. (2012)	stochastic volatility*	F94	Cohen, Malloy, and Pomorski (2012)	opportunistic buy	C181
Chen and Petkova (2012)	average variance of equity returns	F95		opportunistic sell	C182
Eiling (2013)	income growth goods industries	F96	Hirshleifer, Hsu, and Li (2012)	innovative efficiency	C183
	income growth for manufacturing	F97	Li (2012)	abnormal operating cash flows	C184
	income growth for distributive	F98		abnormal production costs	C185
	income growth for service*	F99	Prakash and Sinha (2012)	deferred revenues	C186
	income growth for government*	F100	Price et al. (2012)	earnings conference calls	C187
Boguth and Kuehn (2012)	consumption volatility	F101	So (2012)	earnings forecast optimism	C188
Chang, Christoffersen, and Jacobs (2012)	market skewness	F102	Boons, De Roon, and Szymanowska (2012)	commodity index	F109
Viale, Garcia-Feijoo, and Giannetti (2012)	learning*	F103	Moskowitz, Ooi, and Pedersen (2012)	time-series momentum	C189
	Knightian uncertainty	F104	Koijen et al. (2012)	carry	C190
Bali and Zhou (2012)	market uncertainty	F105	Burlacu et al. (2012)	expected return proxy	C191
[14]Gómez, Priestley, and Zapatero (2012)	labor income‡		Beneish, Lee, and Nichols (2012)	fraud probability	C192
Van Binsbergen (2012)	product price change	C165			

(*continued*)

Table 6
Continued

Reference	Factor	#	Reference	Factor	#
Brennan et al. (2012)	buy orders	C193	Frazzini and Pedersen (2013)	betting-against-beta	C198
	sell orders	C194	Valta (2013)	secured debt	C199
Doskov, Pekkala, and Ribeiro (2013)	expected dividend level	F110		convertible debt	C200
	expected dividend growth	F111		convertible debt indicator	C201
Cohen, Diether, and Malloy (2013)	firm's ability to innovate	C195	Akbas et al. (2013)	cross-sectional pricing inefficiency	F112
Larcker, So, and Wang (2013)	board centrality	C196	Chordia, Subrahmanyam, and Tong (2013)	attenuated returns	C202
Novy-Marx (2013)	gross profitability	C197	Brennan, Huh, and Subrahmanyam (2013)	bad private information	C203
			Han and Zhou (2013)	trend signal	F113

Notes to Table: T, theoretical; F, common factors; C, characteristics. An augmented version (which includes full citations, as well as hyperlinks to each of the cited articles) of this table is available for download and resorting. See http://faculty.fuqua.duke.edu/čharvey/Factor-List.xlsx. Many of the working papers we cite have been published, but because our method depends on the point in time, we cite only the working paper version.

This table contains a summary of risk factors that explain the cross-section of expected returns.

*, insignificant; †, duplicated; ‡, missing p-value.

1: Black, Jensen, and Scholes (1972) first tested the market factor. However, they focus on industry portfolios and thus present a less powerful test compared to Fama and MacBeth (1973). We therefore use the test statistics in Fama and MacBeth (1973) for the market factor.

2: No p-values reported for their factors constructed from principal component analysis.

3: Fama and French (1992) create zero-investment portfolios to test size and book-to-market effects. This is different from the testing approach in Banz (1981). We therefore count Fama and French's (1992) test on size effect as a separate one.

4: No p-values reported for their high order equity index return factors.

5: No p-values reported for their eight risk factors that explain international equity returns.

6: No p-values reported for his return factors.

7: No p-values reported for their five hedge fund style return factors.

8: Vanden (2004) reports a t-statistic for each Fama-French 25 size and book-to-market sorted stock portfolios. We average these 25 t-statistics.

9: Acharya and Pedersen (2005) consider the illiquidity measure in Amihud (2002). This is different from the liquidity measure in Pastor and Stambaugh (2003). We therefore count their factor as a separate one.

10: No p-values reported for the interactions between market return and option returns.

11: No p-values reported for their comoment betas.

12: No p-values reported for his distress tracking factor.

13: Gómez, Priestley, and Zapatero (2012) study census division level labor income. However, most of the division level labor income have a nonsignificant t-statistic. We do not count their factors.

14: No p-values reported for their factors estimated from the long-run risk model.

15: The paper is replaced by Hou, Xue, and Zhang (2014).

initiatives. The multiple testing framework detailed in our paper is true to this advice.

Our research quantifies the warnings of both Fama (1991) and Schwert (2003). We attempt to navigate the zoo and establish new benchmarks to guide empirical asset pricing tests.

Appendix A Multiple Testing When the Number of Tests (M) Is Unknown

The empirical difficulty in applying standard p-value adjustments is that we do not observe factors that have been tried, found to be insignificant and then discarded. We attempt to overcome this difficulty using a simulation framework. The idea is first to simulate the empirical distribution of p-values for all experiments (published and unpublished) and then to adjust the p-values based on these simulated samples.

First, we assume the test statistic (t-statistic, for instance) for any experiment follows a certain distribution D (e.g., exponential distribution) and the set of published works is a truncated D distribution. Based on the estimation framework for truncated distributions,[46] we estimate the parameters of distribution D and the total number of trials M. Next, we simulate many sequences of p-values, each corresponding to a plausible set of p-value realizations of all trials. To account for the uncertainty in parameter estimates of D and M, we simulate the p-value sequences based on the distribution of estimated D and M. Finally, for each p-value, we calculate the adjusted p-value based on a sequence of simulated p-values. The median is taken as the final adjusted p-value.

A.1 Using Truncated Exponential Distribution to Model the t-statistic Sample

Truncated distributions have been used to study hidden tests (i.e., publication bias) in medical research.[47] The idea is that studies reporting significant results are more likely to get published. Assuming a threshold significance level or t-statistic, researchers can, to some extent, infer the results of unpublished works and gain an understanding of the overall effect of a drug or treatment. However, in medical research, insignificant results are still viewed as an indispensable part of the overall statistical evidence and are given much more prominence than in the financial economics research. As a result, medical publications are more likely to report insignificant results. This makes applying the truncated distribution framework to medical studies difficult as there is no clear-cut threshold value.[48] In this sense, the truncated distributional framework suits our study better—1.96 is the obvious hurdle that research needs to overcome to be published.

On the other hand, not all tried factors with a t-statistic above 1.96 are reported. In the quantitative asset management industry, significant results are not published—they are considered "trade secrets." For the academic literature, factors with "borderline" t-statistics are difficult to get published. Thus, our sample is likely missing a number of factors that have t-statistics just over 1.96. To make our inference robust, for our baseline result, we assume all tried factors with t-statistics above 2.57 are observed and ignore those with t-statistics in the range of (1.96, 2.57). We experiment with alternative ways to handle t-statistics in this range.

Many distributions can be used to model the t-statistic sample. One restriction that we think any of these distributions should satisfy is the monotonicity of the density curve. Intuitively, it should be easier to find factors with small t-statistics than with large ones.[49] We choose to use the simplest distribution that incorporates this monotonicity condition: the exponential distribution.

---

46 See Heckman (1979) and Greene (2008), Chapter 24).

47 See Begg and Berlin (1988) and Thornton and Lee (2000).

48 When the threshold value is unknown, it must be estimated from the likelihood function. However, such estimation usually incurs large estimation errors.

49 This basic scarcity assumption is also the key ingredient in our model in Section 5.

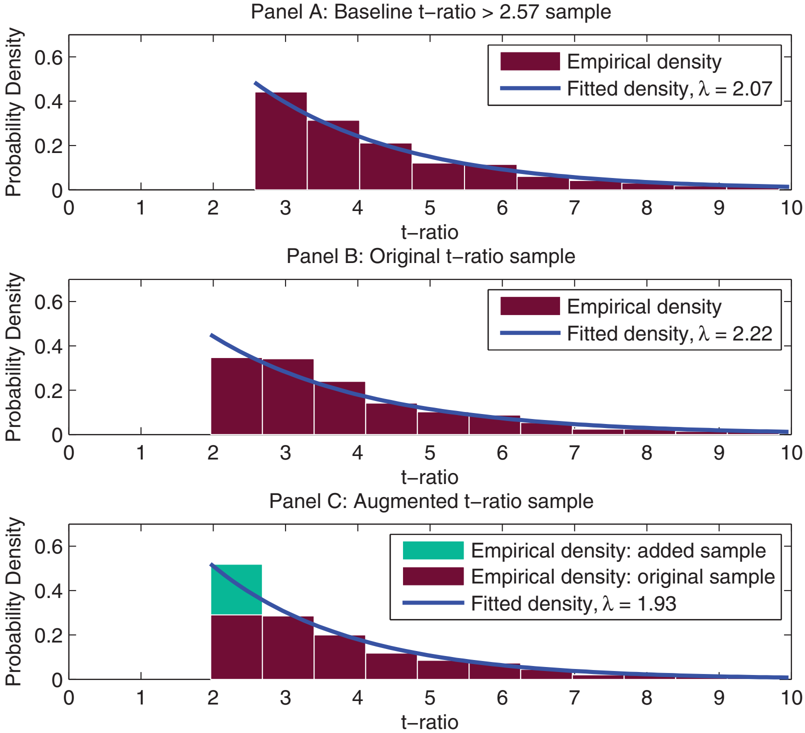

Figure A.1
Density plots for *t*-statistic
Empirical density and fitted exponential density curves based on three different samples. Panel A is based on the baseline sample that includes all t-statistics above 2.57. Panel B is based on the original sample with all t-statistics above 1.96. Panel C is based on the augmented sample that adds the subsample of observations that fall in between 1.96 and 2.57 to the original t-statistic sample. It doubles the number of observations within the range of 1.96 and 2.57 in the original sample. λ is the single parameter for the exponential curve. It gives the population mean for the unrestricted (i.e., nontruncated) distribution.

Panel A of Figure A.1 presents the histogram of the baseline t-statistic sample and the fitted truncated exponential curve.[50] The fitted density closely tracks the histogram and has a population mean of 2.07.[51] Panel B is a histogram of the original t-statistic sample, which, as we discussed before, is likely to underrepresent the sample with a t-statistic in the range of (1.96, 2.57). Panel C is the augmented t-statistic sample with the ad hoc assumption that our sample covers only half of all factors with t-statistics between 1.96 and 2.57. The population mean estimate is 2.22 in panel

[50] There are a few very large t-statistics in our sample. We fit the truncated exponential model without dropping any large t-statistics. In contrast to the usual normal density, exponential distribution is better at modeling extreme observations. In addition, extreme values are pivotal statistics for heavy-tailed distributions and are key for model estimation. While extreme observations are included for model estimation, we exclude them in Figure A.1 to better focus on the main part of the t-statistic range.

[51] Our truncated exponential distribution framework allows a simple analytical estimate for the population mean of the exponential distribution. In particular, let c be the truncation point and the t-statistic sample be $\{t_i\}_{i=1}^{N}$. The mean estimate is given by $\hat{\lambda}=1/(\bar{t}-c)$, where $\bar{t}=(\sum_{i=1}^{N} t_i)/N$ is the sample mean.

B and 1.93 in panel C. As expected, the underrepresentation of relatively small t-statistics results in a higher mean estimate for the t-statistic population. We think the baseline model is the best among all three models as it not only overcomes the missing data problem for the original sample, but also avoids guessing the fraction of missing observations in the 1.96–2.57 range. We use those model estimates for the follow-up analysis.

Using the baseline model, we calculate other interesting population characteristics that are key to multiple hypothesis testing. Assuming independence, we model observed t-statistics as draws from an exponential distribution with mean parameter $\hat{\lambda}$ and a known cutoff point of 2.57. The proportion of unobserved factors is then estimated as

$$P(\text{unobserved}) = \Phi(2.57;\hat{\lambda}) = 1-\exp(-2.57/\hat{\lambda}) = 71.1\%, \tag{A.1}$$

where $\Phi(c;\lambda)$ is the cumulative distribution function evaluated at c for a exponential distribution with mean λ. Our estimates indicate that the mean absolute value of the t-statistic for the underlying factor population is 2.07 and about 71.1% of tried factors are discarded. Given that 238 out of the original 316 factors have a t-statistic exceeding 2.57, the total number of factor tests is estimated to be 824 ($=238/(1-71.1\%)$) and the number of factors with a t-statistic between 1.96 and 2.57 is estimated to be 82.[52] Since our t-statistic sample covers only 57 such factors, roughly 30% (=(82-57)/82) of t-statistics between 1.96 and 2.57 are hidden.

A.2 Simulated Benchmark *t*-statistics under Independence

The truncated exponential distribution framework helps us approximate the distribution of t-statistics for all factors, both published and unpublished. We can then apply the aforementioned adjustment techniques to this distribution to generate new t-statistic benchmarks. However, there are two sources of sampling and estimation uncertainty that affect our results. First, our t-statistic sample may underrepresent all factors with t-statistics exceeding 2.57.[53] Hence, our estimates of the total trials are biased (too low), which affects our calculation of the benchmarks. Second, estimation errors in the truncated exponential distribution can affect our benchmark t-statistics. Although we can approximate the estimation error through the usual asymptotic distribution theory for MLE, it is unclear how this error affects our benchmark t-statistics. This is because t-statistic adjustment procedures usually depend on the entire t-statistic distribution and so standard transformational techniques (e.g., the delta method) do not apply. Moreover, we are not sure whether our sample is large enough to trust the accuracy of asymptotic approximations.

Given these concerns, we propose a four-step simulation framework that incorporates these uncertainties.

Step I Estimate λ and M based on a new *t*-statistic sample with size $r \times R$.

Suppose our current t-statistic sample size is R and it only covers a fraction of $1/r$ of all factors. We sample $r \times R$ t-statistics (with replacement) from the original t-statistic sample. Based on this new t-statistic sample, we apply the above truncated exponential distribution framework to the t-statistics and obtain the parameter estimates λ for the exponential distribution. The truncation probability is calculated as $\hat{P} = \Phi(2.57;\hat{\lambda})$. We can then estimate the total number of trials by

$$\hat{M} = \frac{rR}{1-\hat{P}}.$$

[52] Directly applying our estimate framework to the original sample that includes all t-statistics above 1.96, the estimated total number of factor tests would be 713. Alternatively, assuming our sample only covers half of the factors with t-statistics between 1.96 and 2.57, the estimated number of factors is 971.

[53] This will happen if we miss factors published by the academic literature or we do not have access to the "trade secrets" by industry practitioners.

Table A.1
Benchmark *t*-statistics when *M* is estimated

Sampling ratio	M		Bonferroni		Holm		BHY(1%)		BHY(5%)	
(r)	[10%	90%]	[10%	90%]	[10%	90%]	[10%	90%]	[10%	90%]
1	817		4.01		3.96		3.68		3.17	
	[731	947]	[3.98	4.04]	[3.92	4.00]	[3.63	3.74]	[3.12	3.24]
1.5	1,234		4.11		4.06		3.70		3.20	
	[1,128	1,358]	[4.08	4.13]	[4.03	4.09]	[3.66	3.74]	[3.16	3.24]
2	1,646		4.17		4.13		3.71		3.21	
	[1,531	1,786]	[4.15	4.19]	[4.11	4.15]	[3.67	3.75]	[3.18	3.25]

The estimated total number of factors tried (M) and the benchmark t-statistic percentiles based on a truncated exponential distribution framework. Our estimation is based on the original t-statistic sample truncated at 2.57. The sampling ratio is the assumed ratio of the true population size of t-statistics exceeding 2.57 over our current sample size. Both Bonferroni and Holm have a significance level of 5%.

Step II Calculate the benchmark t-statistic based on a random sample generated from $\hat{\lambda}$ and $\hat{M}$.

Based on the previous step estimate of $\hat{\lambda}$ and $\hat{M}$, we generate a random sample of t-statistics for all tried factors. We then calculate the appropriate benchmark t-statistic based on this generated sample.

Step III Repeat Step II 10,000 times to obtain the median benchmark t-statistic.

We take the median as the final benchmark t-statistic corresponding to the parameter estimate $(\hat{\lambda}, \hat{M})$.

Step IV Repeat Steps I-III 10,000 times to generate a distribution of benchmark t-statistics.

Repeat Steps I-III 10,000 times, each time with a newly generated t-statistic sample as in Step I. For each repetition, we obtain a benchmark t-statistic t_i corresponding to the parameter estimates $(\hat{\lambda}_i, \hat{M}_i)$. In the end, we have a collection of benchmark t-statistics $\{t_i\}_{i=1}^{10000}$.

To see how our procedure works, notice that Steps II and III calculate the theoretical benchmark t-statistic for a t-statistic distribution characterized by $(\hat{\lambda}, \hat{M})$. As a result, the outcome is simply one number and there is no uncertainty around it. Uncertainties are incorporated in Steps I and IV. In particular, by repeatedly sampling from the original t-statistic sample and re-estimating λ and M each time, we take into account the estimation error of the truncated exponential distribution. Also, under the assumption that neglected significant t-statistics follow the empirical distribution of our t-statistic sample, by varying r, we can assess how this underrepresentation of our t-statistic sample affects results.

Table A.1 shows estimates of M and benchmark t-statistics. When $r = 1$, the median estimate for the total number of trials is 817,[54] almost the same as our previous estimate of 820 based on the original sample. Unsurprisingly, the Bonferroni implied benchmark t-statistic (4.01) is larger than 3.78, which is what we obtain when ignoring unpublished works. The Holm implied t-statistic (3.96), while not necessarily increasing in the number of trials, is also higher than before (3.64). The BHY implied t-statistic increases from 3.39 to 3.68 at 1% significance and from 2.78 to 3.18 at 5% significance. As r increases, the sample size M and the benchmark t-statistics for all four

54 Our previous estimate of 820 is a one-shot estimate based on the truncated sample. The results in Table A.1 are based on repeated estimates based on resampled data: we resample many times, and 817 is the median of all these estimates. It is close to the one-shot estimate.

types of adjustments increase. When r doubles, the estimate of M also approximately doubles and the Bonferroni and Holm implied t-statistics increase by about 0.2, whereas the BHY implied t-statistics increase by around 0.03 (under both significance levels).

Appendix B A Bayesian Approach to Multiple Tests

The following framework is adopted from Scott and Berger (2006) and highlights the key issues in Bayesian multiple hypothesis testing.[55] More sophisticated generalizations modify the basic model but are unlikely to change the fundamental hierarchical testing structure.[56] We use this framework to explain the pros and cons of performing multiple testing in a Bayesian framework.

The hierarchical model is as follows:

H1. $(X_i|\mu_i,\sigma^2,\gamma_i)\overset{iid}{\sim}N(\gamma_i\mu_i,\sigma^2)$,

H2. $\mu_i|\tau^2\overset{iid}{\sim}N(0,\tau^2), \gamma_i|p_0\overset{iid}{\sim}Ber(1-p_0)$,

H3. $(\tau^2,\sigma^2)\sim\pi_1(\tau^2,\sigma^2), p_0\sim\pi_2(p_0)$.

We explain each step and the notation in detail

H1. X_i denotes the average return generated from a long-short trading strategy based on a certain factor; μ_i is the unknown mean return; σ^2 is the common variance for returns; and γ_i is an indicator function, with $\gamma_i=0$ indicating a zero factor mean. γ_i is the counterpart of the reject/accept decision in the usual (frequentists') hypothesis testing framework.

H1 therefore says that factor returns are independent conditional on mean $\gamma_i\mu_i$ and common variance σ^2, with $\gamma_i=0$ indicating that the factor is spurious. The common variance assumption may look restrictive, but we can always scale factor returns by changing the dollar investment in the long-short strategy. The crucial assumption is conditional independence of average strategy returns. A certain form of conditional independence is unavoidable for Bayesian hierarchical modeling[57]—probably unrealistic for our application. We can easily think of scenarios in which average returns of different strategies are correlated, even when population means are known. For example, it is well known that two of the most popular factors, the Fama and French (1992) HML and SMB, are correlated.

H2. The first-step population parameters μ_i's and γ_i's are assumed to be generated from two other parametric distributions: μ_i's are independently generated from a normal distribution, and γ_i's are simply generated from a Bernoulli distribution, that is, $\gamma_i=0$ with probability p_0.

The normality assumption for the μ_i's requires the reported X_i's to randomly represent either long/short or short/long strategy returns. If researchers have a tendency to report

[55] We choose to present the full Bayes' approach. An alternative approach—the empirical Bayes' approach—is closely related to the BHY method that controls the false-discovery rate (FDR). See Storey (2003) and Efron and Tibshirani (2002) for the empirical Bayes' interpretation of FDR. For details on the empirical Bayes' method, see Efron et al. (2001), and Efron (2004), 2008). For an in-depth investigation of the differences between the full Bayes' and the empirical-Bayes' approach, see Scott and Berger (2010). For an application of the empirical-Bayes' method in finance, see Markowitz and Xu (1994).

[56] See Meng and Dempster (1987) and Whittemore (2007) for more works on the Bayesian approach in hypothesis testing.

[57] Conditional independence is crucial for the Bayesian framework and the construction of posterior likelihoods. Although it can be extended to incorporate special dependence structures, there is no consensus on how to systematically handle dependence. See Brown et al. (2014) for a discussion of independence in Bayesian multiple testing. They also propose a spatial dependence structure into a Bayesian testing framework.

positive abnormal returns, we need to randomly assign to these returns plus/minus signs. The normality assumptions in both H1 and H2 are important as they are necessary to guarantee the properness of the posterior distributions.

H3. Finally, the two variance variables τ^2 and σ^2 follow a joint prior distribution π_1 and the probability p_0 follows a prior distribution π_2.

Objective or "neutral" priors for π_1 and π_2 can be specified as:

$$\pi_1(\tau^2,\sigma^2) \quad \propto \quad (\tau^2+\sigma^2)^{-2},$$

$$\pi_2(p_0) \quad = \quad \text{Uniform}(0,1).$$

Under this framework, the joint conditional likelihood function for X_i's is simply a product of individual normal likelihood functions and the posterior probability that $\gamma_i = 1$ (discovery) can be calculated by applying Bayes' law. When the number of trials is large, to calculate the posterior probability, we need efficient methods, such as importance sampling, which involves high-dimensional integrals.

One benefit of a Bayesian framework for multiple testing is that the multiplicity penalty term is already embedded. In the frequentists' framework, this is done by introducing FWER or FDR. In a Bayesian framework, the so-called "Ockham's razor effect"[58] automatically adjusts the posterior probabilities when more factors are simultaneously tested.[59] Simulation studies in Scott and Berger (2006) show how the discovery probabilities for a few initial signals increase when more noise is added to the original sample.

However, there are several shortcomings for the Bayesian approach. Some of them are specific to the context of our application and the others are generic to the Bayesian multiple testing framework.

At least two issues arise when applying the Bayesian approach to our factor selection problem. First, we do not observe all tried factors. While we back out the distribution of hidden factors parametrically under the frequentist framework, it is not clear how the missing data and the multiple testing problems can be simultaneously solved under the Bayesian framework. Second, the hierarchical testing framework may be overly restrictive. Both independence and normality assumptions can have a large impact on the posterior distributions. Although normality can be somewhat relaxed by using alternative distributions, the scope of alternative distributions is limited as there are only a few distributions that can guarantee the properness of the posterior distributions. Independence, as we previously discussed, is likely to be violated in our context. In contrast, the three adjustment procedures under the frequentists' framework are able to handle complex data structures since they rely on only fundamental probability inequalities to restrict their objective function—the type I error rate.

There are a few general concerns about the Bayesian multiple testing framework. First, it is not clear what to do after obtaining the posterior probabilities for individual hypotheses. Presumably, we should find a cutoff probability P and reject all hypotheses that have a posterior discovery probability larger than P. But then we return to the initial problem of finding an appropriate cutoff p-value, which is not a clear task. Scott and Berger (2006) suggest a decision-theoretic approach that chooses the cutoff P by minimizing a loss function. The parameters of the loss function, however, are again subjective. Second, the Bayesian posterior distributions are computationally challenging. We document 300 factors, but there are potentially many more if missing factors are taken into account. When M gets large, importance sampling is a necessity. However, results of importance sampling rely on simulations and subjective choices of the centers of the probability distributions

58 See Jefferys and Berger (1992).

59 Intuitively, more complex models are penalized because extra parameters involve additional sources of uncertainty. Simplicity is rewarded in a Bayesian framework as simple models produce sharp predictions. See the discussions in Scott (2009).

for random variables. Consequently, two researchers trying to calculate the same quantity might obtain very different results. Moreover, in multiple testing, the curse of dimensionality generates additional risks for Bayesian statistical inference.[60] These technical issues create additional hurdles for the application of the Bayesian approach.

References

Abarbanell, J. S., and B. J. Bushee. 1998. Abnormal returns to a fundamental analysis strategy. *Accounting Review* 73:19–45.

Acharya, V. V., and L. H. Pedersen. 2005. Asset pricing with liquidity risk. *Journal of Financial Economics* 77:375–410.

Ackert, L. F., and G. Athanassakos. 1997. Prior uncertainty, analyst bias, and subsequent abnormal returns. *Journal of Financial Research* 20:263–73.

Adler, M., and B. Dumas. 1983. International portfolio choice and corporation finance: A synthesis. *Journal of Finance* 38:925–84.

Adrian, T., and J. Rosenberg. 2008. Stock returns and volatility: Pricing the short-run and long-run components of market risk. *Journal of Finance* 63:2997–3030.

Adrian, T., E. Etula, and T. Muir. 2012. Financial intermediaries and the cross-section of asset returns. *Journal of Finance*, Forthcoming.

Ahn, S. C., A. R. Horenstein, and N. Wang. 2012. Determining rank of the beta matrix of a linear asset pricing model. Working Paper.

Akbas, F., W. J. Armstrong, and R. Petkova. 2011. The volatility of liquidity and expected stock returns. Working Paper.

Akbas, F., W. J. Armstrong, S. Sorescu, and A. Subrahmanyam. 2013. Time varying market efficiency in the cross-section of expected stock returns. Working Paper.

Ali, A., L. Hwang, and M. A. Trombley. 2003. Arbitrage risk and the book-to-market anomaly. *Journal of Financial Economics* 69:355–73.

Almeida, H., and M. Campello. 2007. Financial constraints, asset tangibility, and corporate investment. *Review of Financial Studies* 20:1429–60.

Amaya, D., P. Christoffersen, K. Jacobs, and A. Vasquez. 2011. Do realized skewness and kurtosis predict the cross-section of equity returns? Working Paper.

Amihud, Y. 2002. Illiquidity and stock returns: cross-section and time-series effects. *Journal of Financial Markets* 5:31–56.

Amihud, Y., and H. Mendelson. 1986. Asset pricing and the bid-ask spread. *Journal of Financial Economics* 17:223–49.

———. 1989. The effects of beta, bid-ask spread, residual risk, and size on stock returns. *Journal of Finance* 44:479–86.

An, J., S. Bhojraj, and D. T. Ng. 2010. Warranted multiples and future returns. *Journal of Accounting, Auditing & Finance* 25:143–69.

Anderson, C. W., and L. Garcia-Feijóo. 2006. Empirical evidence on capital investment, growth options, and security returns. *Journal of Finance* 61:171–94.

Anderson, E. W., E. Ghysels, and J. L. Juergens. 2005. Do heterogeneous beliefs matter for asset pricing? *Review of Financial Studies* 18:875–924.

[60] See Liang and Kelemen (2008) for a discussion of the computational issues in Bayesian multiple testing.

Ang, A., T. G. Bali, and N. Cakici. 2012. The joint cross section of stocks and options. Working Paper.

Ang, A., J. Chen, and Y. Xing. 2006. Downside risk. *Review of Financial Studies* 19:1191–239.

Ang, A., R. J. Hodrick, Y. Xing, and X. Zhang. 2006. The cross-section of volatility and expected returns. *Journal of Finance* 61:259–99.

Arbel, A., S. A. Carvell, and P. Strebel. 1983. Giraffes, institutions and neglected firms. *Financial Analysts Journal* 39:57–63.

Armstrong, C., S. Banerjee, and C. Corona. 2010. Information quality and the cross-section of expected returns. Working Paper.

Asness, C. S., R. B. Porter, and R. Stevens. 2000. Predicting stock returns using industry-relative firm characteristics. Working Paper.

Asness, C. S., T. J. Moskowitz, and L. H. Pedersen. 2013. Value and momentum everywhere. *Journal of Finance* 68:929–85.

Asquith, P., P. A. Pathak, and J. R. Ritter. 2005. Short interest, institutional ownership and stock returns. *Journal of Financial Economics* 78:243–76.

ATLAS Collaboration. 2012. Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Physics Letters B* 716:1–29.

Avramov, D., T. Chordia, G. Jostova, and A. Philipov. 2007. Momentum and credit rating. *Journal of Finance* 62:2503–20.

———. 2009. Dispersion in analysts' earnings forecasts and credit rating. *Journal of Financial Economics* 91:83–101.

Baik, B., and T. S. Ahn. 2007. Changes in order backlog and future returns. *Seoul Journal of Business* 13:105–26.

Bajgrowicz, P., and O. Scaillet. 2012. Technical trading revisited: False discoveries, persistence tests, and transaction costs. *Journal of Financial Economics* 106:473–91.

Bajgrowicz, P., O. Scaillet, and A. Treccani. 2013. Jumps in high-frequency data: Spurious detections, dynamics, and news. Working Paper.

Baker, M., and J. Wurgler. 2006. Investor sentiment and the cross-section of stock returns. *Journal of Finance* 61:1645–80.

Bailey, D. H., and M. López de Prado. 2014. The deflated Sharpe ratio: correcting for selection bias, backtest overfitting, and non-normality. *Journal of Portfolio Management* 40:94–107.

Balachandran, S., and P. Mohanram. 2011. Using residual income to refine the relationship between earnings growth and stock returns. *Review of Accounting Studies* 17:134–65.

Bali, T. G., and A. Hovakimian. 2009. Volatility spreads and expected stock returns. *Management Science* 55:1797–812.

Bali, T. G., and H. Zhou. 2012. Risk, uncertainty, and expected returns. Working Paper.

Bali, T. G., N. Cakici, and R. F. Whitelaw. 2011. Maxing out: Stocks as lotteries and the cross-section of expected returns. *Journal of Financial Economics* 99:427–46.

Baltussen, G., S. Van Bekkum, and B. Van der Grient. 2012. Unknown unknowns: Vol-of-vol and the cross section of stock returns. Working Paper.

Balvers, R. J., and D. Huang. 2007. Productivity-based asset pricing: Theory and evidence. *Journal of Financial Economics* 86:405–45.

Bandyopadhyay, S. P., A. G. Huang, and T. S. Wirjanto. 2010. The accrual volatility anomaly. Working Paper.

Bansal, R., and S. Viswanathan. 1993. No arbitrage and arbitrage pricing: a new approach. *Journal of Finance* 48:1231–62.

Bansal, R., and A. Yaron. 2005. Risks for the long run: a potential resolution of asset pricing puzzles. *Journal of Finance* 59:1481–509.

Bansal, R., R. F. Dittmar, and C. T. Lundblad. 2005. Consumption, dividends, and the cross section of equity returns. *Journal of Finance* 60:1639–72.

Banz, R. W. 1981. The relationship between return and market value of common stocks. *Journal of Financial Economics* 9:3–18.

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

Barber, B. M., T. Odean, and N. Zhu. 2009. Do retail trades move markets? *Review of Financial Studies* 22:152–86.

Barras, L., O. Scaillet, and R. Wermers. 2010. False discoveries in mutual fund performance: Measuring luck in estimated alphas. *Journal of Finance* 65:179–216.

Basu, S. 1977. Investment performance of common stocks in relation to their price-earnings ratios: a test of the efficient market hypothesis. *Journal of Finance* 32:663–82.

———. 1983. The relationship between earnings' yield, market value and return for NYSE common stocks: further evidence. *Journal of Financial Economics* 12:129–56.

Bauman, W. S., and R. Dowen. 1988. Growth projections and common stock returns. *Financial Analyst Journal* 44:79–80.

Bazdresch, S., F. Belo and X. Lin. 2012. Labor hiring, investment, and stock return predictability in the cross section. Working Paper.

Begg, C. B., and J. A. Berlin. 1988. Publication bias: A problem in interpreting medical data. *Journal of the Royal Statistical Society Series A* 151:419–63.

Beneish, M. D. 1997. Detecting GAAP violation: Implications for assessing earnings management among firms with extreme financial performance. *Journal of Accounting and Public Policy* 16:271–309.

Beneish, M. D., C. M. C. Lee, and D. C. Nichols. 2012. Fraud detection and expected returns. Working Paper.

Benjamini, Y., and Y. Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B* 57:289–300.

Benjamini, Y., and W. Liu. 1999. A step-down multiple hypotheses testing procedure that controls the false discovery rate under independence. *Journal of Statistical Planning and Inference* 82:163–70.

Benjamini, Y., and D. Yekutieli. 2001. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics* 29:1165–88.

Berkman, H., B. Jacobsen, and J. B. Lee. 2011. Time-varying rare disaster risk and stock returns. *Journal of Financial Economics* 101:313–32.

Bhandari, L. C. 1988. Debt/equity ratio and expected common stock returns: Empirical evidence. *Journal of Finance* 43:507–28.

Black, F. 1972. Capital market equilibrium with restricted borrowing. *Journal of Business* 45:444–54.

Black, F., M. C. Jensen, and M. Scholes. 1972. The capital asset pricing model: Some empirical tests. In *Studies in the theory of capital markets,* ed. Michael Jensen, 79–121. New York: Praeger.

Boguth, O., and L. A. Kuehn. 2012. Consumption volatility risk. *Journal of Finance*, Forthcoming.

Boons, M., F. De Roon, and M. Szymanowska. 2012. The stock market price of commodity risk. Working Paper.

Bossaerts, P., and R. M. Dammon. 1994. Tax-induced intertemporal restrictions on security returns. *Journal of Finance* 49:1347–71.

Botosan, C. A. 1997. Disclosure level and the cost of equity capital. *Accounting Review* 72:323–49.

Boudoukh, J., R. Michaely, M. Richardson and M. R. Roberts. 2007. On the importance of measuring payout yield: Implications for empirical asset pricing. *Journal of Finance* 62:877–915.

Boyer, B., T. Mitton, and K. Vorkink. 2010. Expected idiosyncratic skewness. *Review of Financial Studies* 23:169–202.

Bradshaw, M. T., S. A. Richardson, and R. G. Sloan. 2006. The relation between corporate financing activities, analysts' forecasts and stock returns. *Journal of Accounting and Economics* 42:53–85.

Brammer, S., C. Brooks, and S. Pavelin. 2006. Corporate social performance and stock returns: UK evidence from disaggregate measures. *Financial Management* 35:97–116.

Brandt, M., R. Kishore, P. Santa-Clara, and M. Venkatachalam. 2008. Earnings announcements are full of surprises. Working Paper.

Breeden, D. T. 1979. An intertemporal asset pricing model with stochastic consumption and investment opportunities. *Journal of Financial Economics* 7:265–96.

Breeden, D. T., M. R. Gibbons, and R. H. Litzenberger. 1989. Empirical test of the consumption-oriented CAPM. *Journal of Finance* 44:231–62.

Brennan, M. J., T. Chordia, and A. Subrahmanyam. 1998. Alternative factor specifications, security characteristics, and the cross-section of expected stock returns. *Journal of Financial Economics* 49: 345–73.

Brennan, M. J., T. Chordia, A. Subrahmanyam, and Q. Tong. 2012. Sell-order liquidity and the cross-section of expected stock returns. *Journal of Financial Economics* 105:523–41.

Brennan, M. J., S. Huh, and A. Subrahmanyam. 2013. The pricing of good and bad private information in the cross-section of expected stock returns. Working Paper.

Brennan, M. J., and F. Li. 2008. Agency and asset pricing. Working Paper.

Brennan, M. J., and A. Subrahmanyam. 1996. Market microstructure and asset pricing: On the compensation for illiquidity in stock returns. *Journal of Financial Economics* 41:441–64.

Brennan, M. J., A. W. Wang, and Y. Xia. 2004. Estimation and test of a simple model of intertemporal capital asset pricing. *Journal of Finance* 59:1743–76.

Brown, D. A., N. A. Lazar, G. S. Datta, W. Jang, J. E. McDowell. 2014. Incorporating spatial dependence into Bayesian multiple testing of statistical parametric maps in functional neuroimaging. *NeuroImage* 84: 97–112.

Brown, D. P., and B. Rowe. 2007. The productivity premium in equity returns. Working Paper.

Burlacu, R., P. Fontaine, S. Jimenez-Garces, and M. S. Seasholes. 2012. Risk and the cross section of stock returns. *Journal of Financial Economics* 105:511–22.

Callen, J. L., M. Khan, and H. Lu. 2013. Accounting quality, stock price delay, and future stock returns. *Contemporary Accounting Research* 30:269–95.

Callen, J. L., and M. R. Lyle. 2011. The term structure of implied costs of equity capital. Working Paper.

Campbell, J. Y. 1996. Understanding risk and return. *Journal of Political Economy* 104:298–345.

Campbell, J. Y., S. Giglio, C. Polk, and R. Turley. 2012. An intertemporal CAPM with stochastic volatility. Working Paper.

Campbell, J. Y., J. Hilscher, and J. Szilagyi. 2008. In search of distress risk. *Journal of Finance* 63:2899–939.

Campbell, J. Y., and T. Vuolteenaho. 2004. Bad beta, good beta. *American Economic Review* 94:1249–75.

Cao, X., and Y. Xu. 2010. Long-run idiosyncratic volatilities and cross-sectional stock returns. Working Paper.

Cao, C., Y. Chen, B. Liang, and A. W. Lo. 2013. Can hedge funds time market liquidity? *Journal of Financial Economics* 109:493–516.

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

Cen, L., K. C. J. Wei, and J. Zhang. 2006. Forecasted earnings per share and the cross section of expected stock returns. Working Paper.

Chan, K. C., N. Chen, and D. A. Hsieh. 1985. An exploratory investigation of the firm size effect. *Journal of Financial Economics* 14:451–71.

Chan, K. C., S. Foresi, and L. H. P. Lang. 1996. Does money explain asset returns? Theory and empirical analysis. *Journal of Finance* 51:345–61.

Chandrashekar, S., and R. K. S. Rao. 2009. The productivity of corporate cash holdings and the cross-section of expected stock returns. Working Paper.

Chang, B. Y., P. Christoffersen, and K. Jacobs. 2012. Market skewness risk and the cross section of stock returns. *Journal of Financial Economics*, Forthcoming.

Chapman, D. A. 1997. Approximating the asset pricing kernel. *Journal of Finance* 52:1383–410.

Chemmanur, T. J., and A. Yan. 2009. Advertising, attention, and stock returns. Working Paper.

Chen, J., H. Hong, and J. C. Stein. 2002. Breadth of ownership and stock returns. *Journal of Financial Economics* 66:171–205.

Chen, H., M. Kacperczyk, and H. Ortiz-Molina. 2011. Labor unions, operating flexibility, and the cost of equity. *Journal of Financial and Quantitative Analysis* 46:25–58.

Chen, L., R. Novy-Marx, and L. Zhang. 2011. An alternative three-factor model. Working Paper.

Chen, Z., and R. Petkova. 2012. Does idiosyncratic volatility proxy for risk exposure? *Review of Financial Studies* 25:2745–87.

Chen, N. F., R. Roll, and S. A. Ross. 1986. Economic forces and the stock market. *Journal of Business* 59:383–403.

Chen, Z., and I. A. Strebulaev. 2012. Contingent-claim-based expected stock returns. Working Paper.

Chopra, N., J. Lakonishok, and J. R. Ritter. 1992. Measuring abnormal performance: do stocks overreact? *Journal of Financial Economics* 31:235–68.

Chordia, T., S. W. Huh, and A. Subrahmanyam. 2009. Theory-based illiquidity and asset pricing. *Review of Financial Studies* 22:3629–68.

Chordia, T., and L. Shivakumar. 2006. Earnings and price momentum. *Journal of Financial Economics* 80: 627–56.

Chordia, T., A. Subrahmanyam, and V. R. Anshuman. 2001. Trading activity and expected stock returns. *Journal of Financial Economics* 59:3–32.

Chordia, T., A. Subrahmanyam, and Q. Tong. 2013. Have capital market anomalies attenuated in the recent era of high liquidity and trading activity? Working Paper.

Chung, Y. P., H. Johnson, and M. J. Schill. 2006. Asset pricing when returns are nonnormal: Fama-French factors versus higher-order systematic comoments. *Journal of Business* 79:923–40.

CMS Collaboration. 2012. Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Physics Letters B* 716:30–61.

Cochrane, J. H. 1991. Production-based asset pricing and the link between stock returns and economic fluctuations. *Journal of Finance* 46:209–37.

———. 1996. A cross-sectional test of an investment-based asset pricing model. *Journal of Political Economy* 104:572–621.

———. 2011. Presidential Address: Discount Rates. *Journal of Finance* 66:1047–108.

Cohen, L., K. Diether, and C. Malloy. 2013. Misvaluing innovation. *Review of Financial Studies* 26:635–66.

Cohen, L., and A. Frazzini. 2008. Economic links and predictable returns. *Journal of Finance* 63:1977–2011.

Cohen, L., and D. Lou. 2012. Complicated firms. *Journal of Financial Economics* 104:383–400.

Cohen, L., C. Malloy, and L. Pomorski. 2012. Decoding inside information. *Journal of Finance* 67:1009–43.

Cohen, R., C. Polk, and B. Silli. 2009. Best ideas. Working Paper.

Conneely, K. N., and M. Boehnke. 2007. So many correlated tests, so little time! Rapid adjustment of P values for multiple correlated tests. *American Journal of Human Genetics* 81:1158–68.

Conrad, J., M. Cooper, and G. Kaul. 2003. Value versus glamour. *Journal of Finance* 58:1969–96.

Conrad, J., R. F. Dittmar, and E. Ghysels. 2013. Ex ante skewness and expected stock returns. *Journal of Finance* 68:85–124.

Constantinides, G. M. 1982. Intertemporal asset pricing with heterogeneous consumers and without demand aggregation. *Journal of Business* 55:253–67.

———. 1986. Capital market equilibrium with transaction costs. *Journal of Political Economy* 94:842–62.

Cooper, M. J., and H. Gulen. 2006. Is time-series-based predictability evident in real time? *Journal of Business* 79:1263–92.

Cooper, M. J., H. Gulen, and A. V. Ovtchinnikov. 2010. Corporate political contributions and stock returns. *Journal of Finance* 65:687–724.

Cooper, M. J., H. Gulen, and M. J. Schill. 2008. Asset growth and the cross-section of stock returns. *Journal of Finance* 63:1609–51.

Cox, J. C., J. E. Ingersoll, Jr., and S. A. Ross. 1985. An intertemporal general equilibrium model of asset pricing. *Econometrica* 53:363–84.

Cremers, K. J. M., M. Halling, and D. Weinbaum. 2010. In search of aggregate jump and volatility risk in the cross-section of stock returns. Working Paper.

Cremers, K. J. M., and V. B. Nair. 2005. Governance mechanisms and equity prices. *Journal of Finance* 60:2859–94.

Cremers, K. J. M., V. B. Nair, and K. John. 2009. Takeovers and the cross-section of returns. *Review of Financial Studies* 22:1409–45.

Da, Z. 2009. Cash flow, consumption risk, and the cross-section of stock returns. *Journal of Finance* 64:923–56.

Da, Z., Q. Liu, and E. Schaumburg. 2011. Decomposing short-term return reversal. Working Paper.

Da, Z., and E. Schaumburg. 2011. Relative valuation and analyst target price forecasts. *Journal of Financial Markets* 14:161–92.

Da, Z., and M. C. Warachka. 2009. Cash flow risk, systematic earnings revisions, and the cross-section of stock returns. *Journal of Financial Economics* 94:448–68.

———. 2009. Long-term earnings growth forecasts, limited attention, and return predictability. Working Paper.

Daniel, K., and S. Titman. 1997. Evidence on the characteristics of cross sectional variation in stock returns. *Journal of Finance* 52:1–33.

———. 2006. Market reactions to tangible and intangible information. *Journal of Finance* 61:1605–43.

———. 2012. Testing factor-model explanations of market anomalies. *Critical Finance Review* 1:103–39.

Datar, V. T., N. Y. Naik, and R. Radcliffe. 1998. Liquidity and stock returns: An alternative test. *Journal of Financial Markets* 1:203–19.

De Bondt, W. F. M., and R. Thaler. 1985. Does the stock market overreact? *Journal of Finance* 40:793–805.

Dichev, I. D. 1998. Is the risk of bankruptcy a systematic risk? *Journal of Finance* 53:1131–47.

Dichev, I. D., and J. D. Piotroski. 2001. The long-run stock returns following bond ratings changes. *Journal of Finance* 56:173–203.

Diether, K. B., C. J. Malloy, and A. Scherbina. 2002. Differences of opinions and the cross section of stock returns. *Journal of Finance* 57:2113–41.

Dittmar, R. F. 2002. Nonlinear pricing kernels, kurtosis preference, and evidence from the cross section of equity returns. *Journal of Finance* 57:369–403.

Donangelo, A. 2012. Labor mobility: Implications for Asset Pricing. Working Paper.

Doskov, N., T. Pekkala, and R. M. Ribeiro. 2013. Tradable macro risk factors and the cross-section of stock returns. Working Paper.

Douglas, G. W. 1967. Risk in the equity markets: An empirical appraisal of market efficiency. *Yale Economic Essays* 9:3–48.

Doran, J. S., A. Fodor, and D. R. Peterson. 2007. Insiders versus outsiders with employee stock options: Who knows best about future firm risk and implications for stock returns. Working Paper.

Doyle, J. T., R. J. Lundholm, and M. T. Soliman. 2003. The predictive value of expenses excluded from pro forma earnings. *Review of Accounting Studies* 8:145–74.

Drake, M. S., L. Rees, and E. P. Swanson. 2011. Should investors follow the prophets or the bears? Evidence on the use of public information by analysts and short sellers. *Accounting Review* 86: 101–30.

Dudoit, S., and M. J. van der Laan. 2008. *Multiple testing procedures with applications to Genomics.* Springer Series in Statistics. New York.

Easley, D., S. Hvidkjaer, and M. O'Hara. 2002. Is information risk a determinant of asset returns? *Journal of Finance* 57:2185–221.

———. 2010. Factoring information into returns. *Journal of Financial and Quantitative Analysis* 45:293–309.

Eberhart, A. C., W. F. Maxwell, and A. R. Siddique. 2004. An examination of long-term abnormal stock returns and operating performance following R&D increases. *Journal of Finance* 59:623–50.

Edelen, R. M., O. S. Ince, and G. B. Kadlec. 2014. Institutional investors and stock return anomalies. Working Paper.

Edmans, A. 2011. Does the stock market fully value intangibles? Employee satisfaction and equity prices. *Journal of Financial Economics* 101:621–40.

Efron, B. 1979. Bootstrap methods: Another look at the jackknife. *Annals of Statistics* 7:1–26.

———. 2004. Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *Journal of the American Statistical Association* 99:96–104.

———. 2008. Microarrays, empirical Bayes, and the two-groups model. *Statistical Science* 23:1–22.

Efron, B., and R. Tibshirani. 2002. Empirical Bayes methods and false discovery rates for microarrays. *Genetic Epidemiology* 23:70–86.

Efron, B., R. Tibshirani, J. D. Storey, and V. Tusher. 2001. Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association* 96:1151–60.

Eiling, E. 2013. Industry-specific human capital, idiosyncratic risk, and the cross-section of expected stock returns. *Journal of Finance* 68:43–84.

Eisfeldt, A. L., and D. Papanikolaou. 2011. Organization capital and the cross-section of expected returns. Working Paper.

Elgers, P. T., M. H. Lo, and R. J. Pfeiffer, Jr., 2001. Delayed security price adjustments to financial analysts' forecasts of annual earnings. *Accounting Review* 76:613–32.

Elton, E. J., M. J. Gruber, and C. R. Blake. 1995. Fundamental economic variables, expected returns, and bond fund performance. *Journal of Finance* 50:1229–56.

Elton, E. J., M. J. Gruber, S. Das, and M. Hlavka. 1993. Efficiency with costly information: A reinterpretation of evidence from managed portfolios. *Review of Financial Studies* 6:1–22.

Erb, C. B., C. R. Harvey, and T. E. Viskanta. 1996. Expected returns and volatility in 135 countries. *Journal of Portfolio Management* 22:46–58.

Fabozzi, F. J., K. C. Ma, and B. J. Oliphant. 2008. Sin stock returns. *Journal of Portfolio Management* 35:82–94.

Fairfield, P. M., J. S. Whisenant, and T. L. Yohn. 2003. Accrued earnings and growth: implications for future profitability and market mispricing. *Accounting Review* 78:353–71.

Fama, E. F. 1991. Efficient capital markets: II. *Journal of Finance* 46:1575–617.

Fama, E. F., and K. R. French. 1992. The cross-section of expected stock returns. *Journal of Finance* 47:427–65.

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

———. 2006. Profitability, investment and average returns. *Journal of Financial Economics* 82:491–518.

———. 2010. Luck versus skill in the cross section of mutual fund returns. *Journal of Finance* 65:1915–47.

Fama, E. F., and J. D. MacBeth. 1973. Risk, return, and equilibrium: Empirical tests. *Journal of Political Economy* 81:607–36.

Fang, L., and J. Peress. 2009. Media coverage and the cross-section of stock returns. *Journal of Finance* 64:2023–52.

Farcomeni, A. 2007. A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. *Statistical Methods in Medical Research* 17:347–88.

Ferson, W. E., and Y. Chen. 2013. How many good and bad fund managers are there, really? Working Paper.

Ferson, W. E. and C. R. 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.

———. 1994. Sources of risk and expected returns in global equity markets. *Journal of Banking and Finance* 18:775–803.

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

Ferson, W. E., S. Nallareddy, and B. Xie. 2012. The "out-of-sample" performance of long run risk models. *Journal of Financial Economics*, Forthcoming.

Figlewski, S. 1981. The informational effects of restrictions on short sales: some empirical evidence. *Journal of Financial and Quantitative Analysis* 16:463–76.

Fogler, H. R., K. John, and J. Tipton. 1981. Three factors, interest rate differentials and stock groups. *Journal of Finance* 36:323–35.

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

Frank, M. Z., and V. K. Goyal. 2009. Capital structure decisions: Which factors are reliably important? *Financial Management* 38:1–37.

Frankel, R., and C. M. C. Lee. 1998. Accounting valuation, market expectation, and cross-sectional stock returns. *Journal of Accounting and Economics* 25:283–319.

Franzoni, F., and J. M. Marin. 2006. Pension plan funding and stock market efficiency. *Journal of Finance* 61:921–56.

Frazzini, A., and L. H. Pedersen. 2013. Betting against beta. Working Paper.

Friewald, N., C. Wagner, and J. Zechner. 2012. The cross-section of credit risk premia and equity returns. Working Paper.

Fu, F. 2009. Idiosyncratic risk and the cross-section of expected stock returns. *Journal of Financial Economics* 91:24–37.

Fung, W., and D. A. Hsieh. 1997. Empirical characteristics of dynamic trading strategies: The case of hedge funds. *Review of Financial Studies* 10:275–302.

———. 2001. The risk in hedge fund strategies: theory and evidence from trend followers. *Review of Financial Studies* 14:313–41.

Garcia, D., and Ø. Norli. 2012. Geographic dispersion and stock returns. *Journal of Financial Economics*, Forthcoming.

Garlappi, L., and H. Yan. 2011. Financial distress and the cross-section of equity returns. *Journal of Finance* 66:789–822.

Garlappi, L., T. Shu, and H. Yan. 2008. Default risk, shareholder advantage, and stock returns. *Review of Financial Studies* 21:2743–78.

Gârleanu, N., L. Kogan, and S. Panageas. 2012. Displacement risk and asset returns. *Journal of Financial Economics* 105:491–510.

Ge, Y., S. Dudoit, and T. P. Speed. 2003. Resampling-based multiple testing for microarray data analysis. *Test* 12:1–77.

George, T. J., and C. Y. Hwang. 2004. The 52-week high and momentum investing. *Journal of Finance* 59:2145–76.

———. 2010. A resolution of the distress risk and leverage puzzles in the cross section of stock returns. *Journal of Financial Economics* 96:56–79.

Gettleman, E., and J. M. Marks. 2006. Acceleration strategies. Working Paper.

Glaeser, E. 2008. Research incentives and empirical methods. In *The Foundations of positive and normative economics: A handbook*, Chapter 13. Oxford: Oxford University Press.

Gokcen, U. 2009. Information revelation and expected stock returns. Working Paper.

Gomes, J. F., A. Yaron, and L. Zhang. 2006. Asset pricing implications of firms' financing constraints. *Review of Financial Studies* 19:1321–56.

Gómez, J. P., R. Priestley, and F. Zapatero. 2012. Labor income, relative wealth concerns, and the cross-section of stock returns. Working Paper.

Gompers, P. A., and A. Metrick. 2001. Institutional investors and equity prices. *Quarterly Journal of Economics* 116:229–59.

Gompers, P. A., J. L. Ishii, and A. Metrick. 2003. Corporate governance and equity prices. *Quarterly Journal of Economics* 118:107–55.

Gourio, F. 2007. Labor leverage, firms' heterogeneous sensitivities to the business cycle, and the cross-section of expected returns. Working Paper.

Gow, I. D., and D. J. Taylor. 2009. Earnings volatility and the cross-section of returns. Working Paper.

Green, J., J. R. M. Hand, and X. F. Zhang. 2013a. The supraview of return predictive signals. *Review of Accounting Studies* 18:692–730.

———. 2013b. The remarkable multidimensionality in the cross section of expected US stock returns. Working Paper.

Greene, W. H. 2008. *Econometric analysis*. Upper Saddle River, NJ: Prentice Hall.

Griffin, J. M., and M. L. Lemmon. 2002. Book-to-market equity, distress risk, and stock returns. *Journal of Finance* 57:2317–36.

Gu, F. 2005. Innovation, future earnings, and market efficiency. *Journal of Accounting, Auditing and Finance* 20:385–418.

Gu, F., and B. Lev. 2011. Overpriced shares, ill-advised acquisitions, and goodwill impairment. *Accounting Review* 86:1995–2022.

Gu, L., Z. Wang, and J. Ye. 2008. Information in order backlog: Change versus level. Working Paper.

Guo, H., and R. Savickas. 2008. Average idiosyncratic volatility in G7 countries. *Review of Financial Studies* 21:1259–96.

Gupta, M. C., and A. R. Ofer. 1975. Investor's expectations of earnings growth, their accuracy and effects on the structure of realized rates of return. *Journal of Finance* 30:509–23.

Hafzalla, N., R. Lundholm, and E. M. Van Winkle. 2011. Percent accruals. *Accounting Review* 86:209–36.

Hahn, J., and H. Lee. 2009. Financial constraints, debt capacity, and the cross-section of stock returns. *Journal of Finance* 64:891–921.

Hameed, A., J. Huang, and G. M. Mian. 2010. Industries and stock return reversals. Working Paper.

Han, B., and Y. Zhou. 2011. Term structure of credit default swap spreads and cross-section of stock returns. Working Paper.

Han, B., H. M. Kang, and E. Eskin. 2009. Rapid and accurate multiple testing correction and power estimation for millions of correlated markers. *PLoS genetics* 5:e1000456.

Han, Y., and G. Zhou. 2013. Trend factor: A new determinant of cross-section stock returns. Working Paper.

Harvey, C. R., and A. Siddique. 2000. Conditional skewness in asset pricing tests. *Journal of Finance* 55:1263–95.

Harvey, C. R., and Y. Liu. 2014a. Multiple testing in economics. Working Paper.

———. 2014b. Evaluating trading strategies. *Journal of Portfolio Management* 40:108–18.

———. 2014c. Lucky factors. Working Paper.

Hawkins, E. H., S. C. Chamberlin, and W. E. Daniel. 1984. Earnings expectations and security prices. *Financial Analysts Journal* 40:24–74.

Head, A., G. Smith, and J. Wilson. 2007. Would a stock by any other ticker smell as sweet? *Quarterly Review of Economics and Finance* 49:551–61.

Heaton, J., and D. Lucas. 2000. Portfolio choice and asset prices: The importance of entrepreneurial risk. *Journal of Finance* 55:1163–98.

Heckerman, D. G. 1972. Portfolio selection and the structure of capital asset prices when relative prices of consumption goods may change. *Journal of Finance* 27:47–60.

Heckman, J. J. 1979. Sample selection bias as a specification error. *Econometrica* 47:153–61.

Hess, D., D. Kreutzmann, and O. Pucker. 2011. Projected earnings accuracy and profitability of stock recommendations. Working Paper.

Hirshleifer, D., P. H. Hsu, and D. Li. 2012. Innovative efficiency and stock returns. *Journal of Financial Economics* 107:632–54.

Hirshleifer, D., and D. Jiang. 2010. A financing-based misvaluation factor and the cross-section of expected returns. *Review of Financial Studies* 23:3401–36.

Hirshleifer, D., K. Kou, S. H. Teoh, and Y. Zhang. 2004. Do investors overvalue firms with bloated balance sheets? *Journal of Accounting and Economics* 38:297–331.

Hochberg, Y. 1988. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* 75:800–02.

Hochberg, Y., and Y. Benjamini. 1990. More powerful procedures for multiple significance testing. *Statistics in Medicine* 9:811–18.

Hochberg, Y., and A. C. Tamhane. 1987. Multiple comparison procedures. John Wiley & Sons.

Holland, B., S. Basu, and F. Sun. 2010. Neglect of multiplicity when testing families of related hypotheses. Working Paper.

Holm, S. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 6:65–70.

Holthausen, R. W., and D. F. Larcker. 1992. The prediction of stock returns using financial statement information. *Journal of Accounting & Economics* 15:373–411.

Hommel, G. 1988. A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika* 75:383–86.

Hou, K., G. A. Karolyi, and B. C. Kho. 2011. What factors drive global stock returns? *Review of Financial Studies* 24:2527–74.

Hou, K., and T. J. Moskowitz. 2005. Market frictions, price delay, and the cross-section of expected returns. *Review of Financial Studies* 18:981–1020.

Hou, K., and D. T. Robinson. 2006. Industry concentration and average stock returns. *Journal of Finance* 61:1927–56.

Hou, K., C. Xue, and L. Zhang. 2014. Digesting anomalies: An investment approach. *Review of Financial Studies*, Forthcoming.

Hu, G. X., J. Pan, and J. Wang. 2012. Noise as information for illiquidity. Working Paper.

Huang, A. G. 2009. The cross section of cashflow volatility and expected stock returns. *Journal of Empirical Finance* 16:409–29.

Huang, W., Q. Liu, S. G. Rhee, and F. Wu. 2012. Extreme downside risk and expected stock returns. *Journal of Banking & Finance* 36:1492–502.

Hvidkjaer, S. 2008. Small trades and the cross-section of stock returns. *Review of Financial Studies* 21:1123–51.

Imrohoroglu, A., and S. Tuzel. 2011. Firm level productivity, risk, and return. Working Paper.

Ioannidis, J. P. A. 2005. Why most published research findings are false? *PLoS Medicine* 2 e124:696–701.

Jacobs, K., and K. Q. Wang. 2004. Idiosyncratic consumption risk and the cross section of asset returns. *Journal of Finance* 59:2211–52.

Jagannathan, R., and Z. Wang. 1996. The conditional CAPM and the cross-section of expected returns. *Journal of Finance* 51:3–53.

Jagannathan, R., and Y. Wang. 2007. Lazy investors, discretionary consumption, and the cross-section of stock returns. *Journal of Finance* 62:1623–61.

Jarrow, R. 1980. Heterogeneous expectations, restrictions on short sales, and equilibrium asset prices. *Journal of Finance* 35:1105–13.

Jefferys, W. H., and J. O. Berger. 1992. Ockham's razor and Bayesian analysis. *American Scientist* 80:64–72.

Jegadeesh, N. 1990. Evidence of predictable behavior of security returns. *Journal of Finance* 45:881–98.

Jegadeesh, N., J. Kim, S. D. Krische, and C. M. C. Lee. 2004. Analyzing the analysts: When do recommendations add value? *Journal of Finance* 59:1083–124.

Jegadeesh, N., and S. Titman. 1993. Returns to buying winners and selling losers: implications for stock market efficiency. *Journal of Finance* 48:65–91.

Jiang, G., C. M. C. Lee, and Y. Zhang. 2005. Information uncertainty and expected returns. *Review of Accounting Studies* 10:185–221.

Jiang, H., and Z. Sun. 2011. Dispersion in beliefs among active mutual funds and the cross-section of stock returns. Working Paper.

Johnson, T. L. and E. C. So. 2012. The option to stock volume ratio and future returns. *Journal of Financial Economics* 106:262–86.

Jones, C. M., and O. A. Lamont. 2002. Short-sale constraints and stock returns. *Journal of Financial Economics* 66:207–39.

Kapadia, N. 2011. Tracking down distress risk. *Journal of Financial Economics* 102:167–82.

Kaplan, S. N., and L. Zingales. 1997. Do investment-cash flow sensitivities provide useful measures of financing constraints? *Quarterly Journal of Economics* 112:169–215.

Karolyi, G. A., and B. C. Kho. 2004. Momentum strategies: some bootstrap tests. *Journal of Empirical Finance* 11:509–36.

Kelly, B., and S. Pruitt. 2011. The three-pass regression filter: A new approach to forecasting using many predictors. Working Paper.

Kim, C. F., C. Pantzalis, and J. C. Park. 2012. Political geography and stock returns: The value and risk implications of proximity to political power. *Journal of Financial Economics* 106:196–228.

Koijen, R. S. J., T. J. Moskowitz, L. H. Pedersen, and E. B. Vrugt. 2012. Carry. Working Paper.

Korajczyk, R. A., and R. Sadka. 2008. Pricing the commonality across alternative measures of liquidity. *Journal of Financial Economics* 87:45–72.

Korniotis, G. M. 2008. Habit formation, incomplete markets, and the significance of regional risk for expected returns. *Review of Financial Studies* 21:2139–72.

Korniotis, G. M., and A. Kumar. 2009. Long Georgia, short Colorado? The geography of return predictability. Working Paper.

Kosowski, R., N. Y. Naik, and M. Teo. 2007. Do hedge funds deliver alpha? A Bayesian and bootstrap analysis. *Journal of Financial Economics* 84:229–64.

Kosowski, R., A. Timmermann, R. Wermers, and H. White. 2006. Can mutual fund “stars” really pick stocks? New evidence from a Bootstrap analysis. *Journal of Finance* 61:2551–95.

Kothari, S. P., and J. B. Warner. 2007, Econometrics of event studies. In: *Handbook of corporate finance: Empirical corporate finance*, Volume I, Ed. B. E. Eckbo, 3–36. Amsterdam: Elsevier.

Kraus, A., and R. H. Litzenberger. 1976. Skewness preference and the valuation of risk assets. *Journal of Finance* 31:1085–100.

Kumar, A., and C. M. C. Lee. 2006. Retail investor sentiment and return comovement. *Journal of Finance* 61:2451–86.

Kumar, P., S. M. Sorescu, R. D. Boehme, and B. R. Danielsen. 2008. Estimation risk, information, and the conditional CAPM: Theory and evidence. *Review of Financial Studies* 21:1037–75.

Kyle, A. S. 1985. Continuous auctions and insider trading. *Econometrica* 53:1315–35.

Lamont, O., C. Polk, and J. Saa-Requejo. 2001. Financial constraints and stock returns. *Review of Financial Studies* 14:529–54.

Landsman, W. R., B. L. Miller, K. Peasnell, and S. Yeh. 2011. Do investors understand really dirty surplus? *Accounting Review* 86:237–58.

La Porta, R. 1996. Expectations and the cross-section of stock returns. *Journal of Finance* 51:1715–42.

Larcker, D. F., E. C. So, and C. C. Y. Wang. 2013. Boardroom centrality and firm performance. *Journal of Accounting and Economics* 55:225–50.

Leamer, E. E. 1978. *Specification searches: Ad hoc inference with nonexperimental data.* New York: John Wiley & Sons.

Lee, C. M. C., and B. Swaminathan. 2000. Price momentum and trading volume. *Journal of Finance* 55:2017–69.

Lehavy, R., and R. G. Sloan. 2008. Investor recognition and stock returns. *Review of Accounting Studies* 13:327–61.

Lehmann, E. L., and J. P. Romano. 2005. Generalizations of the familywise error rate. *Annals of Statistics* 33:1138–54.

Lettau, M., and S. Ludvigson. 2001. Resurrecting the (C)CAPM: A cross-sectional test when risk premia are time-varying. *Journal of Political Economy* 109:1238–87.

Lev, B., D. Nissim, and J. Thomas. 2005. On the informational usefulness of R&D capitalization and amortization. Working Paper.

Lev, B., B. Sarath, and T. Sougiannis. 2005. R&D reporting biases and their consequences. *Contemporary Accounting Research* 22:977–1026.

Lev, B., and T. Sougiannis. 1996. The capitalization, amortization, and value-relevance of R&D. *Journal of Accounting and Economics* 21:107–38.

Lewellen, J., S. Nagel, and J. Shanken. 2010. A skeptical appraisal of asset pricing tests. *Journal of Financial Economics* 96:175–94.

Liang, Y., and A. Kelemen. 2008. Statistical advances and challenges for analyzing correlated high dimensional SNP data in genomic study for complex diseases. *Statistics Surveys* 2:43–60.

Li, D. 2011. Financial constraints, R&D investment, and stock returns. *Review of Financial Studies* 24:2975–3007.

Li, K. K. 2011. How well do investors understand loss persistence? *Review of Accounting Studies* 16:630–67.

Li, Q., M. Vassalou, and Y. Xing. 2006. Sector investment growth rates and the cross section of equity returns. *Journal of Business* 79:1637–65.

Li, S. Z. 2012. Continuous beta, discontinuous beta, and the cross-section of expected stock returns. Working Paper.

Li, X. 2012. Real earnings management and subsequent stock returns. Working Paper.

Lin, D. Y. 2005. An efficient Monte Carlo approach to assessing statistical significance in genomic studies. *Bioinformatics* 21:781–7.

Lintner, J. 1965. Security prices, risk, and maximal gains from diversification. *Journal of Finance* 20:587–615.

Lioui, A., and P. Maio. 2012. Interest rate risk and the cross-section of stock returns. Working Paper.

Litzenberger, R. H., and K. Ramaswamy. 1979. The effect of personal taxes and dividends on capital asset prices: Theory and empirical evidence. *Journal of Financial Economics* 7:163–95.

Liu, W. 2006. A liquidity-augmented capital asset pricing model. *Journal of Financial Economics* 82:631–71.

Liu, Q., Lei L., Bo S., and H. Yan. 2014. A model of anomaly discovery. Working Paper.

Livdan, D., H. Sapriza, and L. Zhang. 2009. Financially constrained stock returns. *Journal of Finance* 64:1827–62.

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

Lo, A. W., and J. Wang. 2006. Trading volume: Implications of an intertemporal capital asset pricing model. *Journal of Finance* 61:2805–40.

Loughran, T., and J. R. Ritter. 1995. The new issues puzzle. *Journal of Finance* 50:23–51.

Loughran, T., and A. M. Vijh. 1997. Do long-term shareholders benefit from corporate acquisitions? *Journal of Finance* 52:1765–90.

Lucas, R. E. Jr., 1978. Asset prices in an exchange economy. *Econometrica* 46:1429–45.

Lustig, H. N., and S. G. Van Nieuwerburgh. 2005. Housing collateral, consumption insurance, and risk premia: An empirical perspective. *Journal of Finance* 60:1167–219.

Lynch, A., and T. Vital-Ahuja. 2012. Can subsample evidence alleviate the data-snooping problem?: A comparison to the maximal R^2 cutoff test. Working Paper.

Malloy, C. J., T. J. Moskowitz, and A. Vissing-Jorgensen. 2009. Long-run stockholder consumption risk and asset returns. *Journal of Finance* 64:2427–79.

Markowitz, H. M., and G. L. Xu. 1994. Data mining corrections. *Journal of Portfolio Management* 21:60–9.

Mayshar, J. 1981. Transaction costs and the pricing of assets. *Journal of Finance* 36:583–97.

McConnell, J. J., and G. C. Sanger. 1984. A trading strategy for new listings on the NYSE. *Financial Analysts Journal* 40:34–8.

McLean, R. D., and J. Pontiff. 2015. Does academic research destroy stock return predictability? *Journal of Finance*, Forthcoming.

Meinshausen, N. 2008. Hierarchical testing of variable importance. *Biometrika* 95:265–78.

Meng, C. Y. K., and A. P. Dempster. 1987. A Bayesian approach to the multiplicity problem for significance testing with binomial data. *Biometrics* 43:301–11.

Menzly, L., and O. Ozbas. 2010. Market segmentation and cross-predictability of returns. *Journal of Finance* 65:1555–80.

Merton, R. C. 1973. An intertemporal capital asset pricing model. *Econometrica* 41:867–87.

Michaely, R., R. H. Thaler, and K. L. Womack. 1995. Price reactions to dividend initiations and omissions: overreaction or drift? *Journal of Finance* 50:573–608.

Mohanram, P. S. 2005. Separating winners from losers among low book-to-market stocks using financial statement analysis. *Review of Accounting Studies* 10:133–70.

Moskowitz, T. J., and M. Grinblatt. 1999. Do industries explain momentum? *Journal of Finance* 54: 1249–90.

Moskowitz, T. J., Y. H. Ooi, and L. H. Pedersen. 2012. Time series momentum. *Journal of Financial Economics* 104:228–50.

Mossin, J. 1966. Equilibrium in a capital asset market. *Econometrica* 34:768–83.

Nagel, S. 2005. Short sales, institutional investors and the cross-section of stock returns. *Journal of Financial Economics* 78:277–309.

Narayanamoorthy, G. 2006. Conservatism and cross-sectional variation in the post-earnings announcement drift. *Journal of Accounting Research* 44:763–89.

Nguyen, G. X., and P. E. Swanson. 2009. Firm characteristics, relative efficiency and equity returns. *Journal of Financial and Quantitative Analysis* 44:213–36.

Novy-Marx, R. 2013. The other side of value: The gross profitability premium. *Journal of Financial Economics* 108:1–28.

———. 2014. Predicting anomaly performance with politics, the weather, global warming, sunspots, and the stars. *Journal of Financial Economics* 112:137–46.

Nyberg, P., and S. Pöyry. 2011. Firm expansion and stock price momentum. Working Paper.

Ofek, E., M. Richardson, and R. F. Whitelaw. 2004. Limited arbitrage and short sales restrictions: evidence from the options markets. *Journal of Financial Economics* 74:305–42.

Oldfield, G. S. Jr., and R. J. Rogalski. 1981. Treasury bill factors and common stock returns. *Journal of Finance* 36:337–50.

Ortiz-Molina, H., and G. M. Phillips. 2011. Real asset liquidity and the cost of capital. Working Paper.

Ou, J. A., and S. H. Penman. 1989. Financial statement analysis and the prediction of stock returns. *Journal of Accounting and Economics* 11:295–329.

Ozoguz, A. 2009. Good times or bad times? Investor's uncertainty and stock returns. *Review of Financial Studies* 22:4377–422.

Palazzo, B. 2012. Cash holdings, risk, and expected returns. *Journal of Financial Economics* 104:162–85.

Papanastasopoulos, G., D. Thomakos, and T. Wang. 2010. The implications of retained and distributed earnings for future profitability and stock returns. *Review of Accounting & Finance* 9:395–423.

Parker, J. A., and C. Julliard. 2005. Consumption risk and the cross section of expected returns. *Journal of Political Economy* 113:185–222.

Pastor, L., and R. F. Stambaugh. 2003. Liquidity risk and expected stock returns. *Journal of Political Economy* 111:643–85.

Patatoukas, P. N. 2011. Customer-base concentration: implications for firm performance and capital markets. Working Paper.

Patton, A. J., and A. Timmermann. 2010. Monotonicity in asset returns: New tests with applications to the term structure, the CAPM, and portfolio sorts. *Journal of Financial Economics* 98:605–25.

Penman, S., and X. Zhang. 2002. Modeling sustainable earnings and P/E ratios with financial statement analysis. Working Paper.

Pesaran, M. H., and A. Timmermann. 2007. Selection of estimation window in the presence of breaks. *Journal of Econometrics* 137:134–61.

Piotroski, J. D. 2000. Value investing: The use of historical financial statement information to separate winners from losers. *Journal of Accounting Research* 38:1–41.

Pontiff, J., and A. Woodgate. 2008. Share issuance and cross-sectional returns. *Journal of Finance* 63:921–45.

Prakash, R., and N. Sinha. 2012. Deferred revenues and the matching of revenues and expenses. *Contemporary Accounting Research* Forthcoming.

Price, S. M., J. S. Doran, D. R. Peterson, and B. A. Bliss. 2012. Earnings conference calls and stock returns: The incremental informativeness of textual tone. *Journal of Banking and Finance* 36:992–1011.

Pukthuanthong, K., and R. Roll. 2014. A protocol for factor identification. Working Paper.

Rajgopal, S., T. Shevlin, and M. Venkatachalam. 2003. Does the stock market fully appreciate the implications of leading indicators for future earnings? Evidence from order backlog. *Review of Accounting Studies* 8:461–92.

Roll, R. 1988. R^2. *Journal of Finance* 43:541–66.

Romano, J. P., A. M. Shaikh, and M. Wolf. 2008. Formalized data snooping based on generalized error rates. *Econometric Theory* 24:404–47.

Rosenthal, R. 1979. The "file drawer problem" and tolerance for null results. *Psychological Bulletin* 86:638–41.

Ross, S. A. 1989. Regression to the max. Working Paper.

Rubinstein, M. E. 1973. The fundamental theorem of parameter-preference security valuation. *Journal of Financial and Quantitative Analysis* 8:61–69.

———. 1974. An aggregation theorem for securities markets. *Journal of Financial Economics* 1:225–44.

Sarkar, S. K. 2002. Some results on false discovery rate in stepwise multiple testing procedure. *Annals of Statistics* 30:239–57.

Sarkar, S. K., and W. Guo. 2009. On a generalized false discovery rate. *Annals of Statistics* 37:1545–65.

Sadka, R. 2006. Momentum and post-earnings-announcement drift anomalies: The role of liquidity risk. *Journal of Financial Economics* 80:309–49.

Saville, D. J. 1990. Multiple comparison procedures: The practical solution. *American Statistician* 44:174–80.

Savov, A. 2011. Asset pricing with garbage. *Journal of Finance* 66:177–201.

Scheffé, H. 1959. *The analysis of variance.* New York: Wiley.

Schweder, T., and E. Spjøtvoll. 1982. Plots of p-values to evaluate many tests simultaneously. *Biometrika* 69:493–502.

Schwert, G. W. 2003. Anomalies and market efficiency. In *Handbook of the economics of finance*. Eds. G. M. Constantinides, M. Harris, and R. Stulz. Amsterdam: Elsevier Science.

Scott, J. G. 2009. Bayesian adjustment for multiplicity. Working Paper.

Scott, J. G., and J. O. Berger. 2006. An exploration of aspects of Bayesian multiple testing. *Journal of Statistical Planning and Inference* 136:2144–62.

———. 2010. Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Annuals of Statistics* 38:2587–619.

Shaffer, J. P. 1995. Multiple hypothesis testing. *Annual Review of Psychology* 46:561–84.

Shanken, J. 1990. Intertemporal asset pricing: An empirical investigation. *Journal of Econometrics* 45:99–120.

Sharpe, W. F. 1964. Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance* 19:425–42.

Shu, T. 2007. Trader composition, price efficiency, and the cross-section of stock returns. Working Paper.

Simes, R. J. 1986. An improved Bonferroni procedure for multiple tests of significance. *Biometrika* 73:751–54.

Simutin, M. 2010. Excess cash and stock returns. *Financial Management* 39:1197–222.

Sloan, R. G. 1996. Do stock prices fully reflect information in accruals and cash flows about future earnings? *Accounting Review* 71:289–315.

So, E. C. 2012. A new approach to predicting analyst forecast errors: Do investors overweight analyst forecasts? Working Paper.

Soliman, M. T. 2008. The use of DuPont analysis by market participants. *Accounting Review* 83:823–53.

Solnik, B. H. 1974. An equilibrium model of the international capital market. *Journal of Economic Theory* 8:500–24.

Spiess, D. K., and J. Affleck-Graves. 1999. Underperformance in long-run stock returns following seasoned equity offerings. *Journal of Financial Economics* 38:243–67.

———. 1999. The long-run performance of stock returns following debt offerings. *Journal of Financial Economics* 54:45–73.

Storey, J. D. 2003. The positive false discovery rate: A Bayesian interpretation and the q-value. *Annals of Statistics* 31:2013–35.

Stulz, R. M. 1981. A model of international asset pricing. *Journal of Financial Economics* 9:383–406.

———. 1986. Asset pricing and expected inflation. *Journal of Finance* 41:209–23.

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 mining: The case of calendar effects in stock returns. *Journal of Econometrics* 105:249–86.

Subrahmanyam, A. 2010. The cross-section of expected stock returns: What have we learnt from the past twenty-five years of research? Working Paper.

Sun, W., and T. T. Cai. 2009. Large-scale multiple testing under dependence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 71:393–424.

Sweeney, R. J., and A. D. Warga. 1986. The pricing of interest-rate risk: evidence from the stock market. *Journal of Finance* 41:393–410.

Teo, M., and S. J. Woo. 2004. Style effects in the cross-section of stock returns. *Journal of Financial Economics* 74:367–98.

Thomas, J., and F. X. Zhang. 2011. Tax expense momentum. *Journal of Accounting Research* 49:791–821.

Thornton, A., and P. Lee. 2000. Publication bias in meta-analysis: its causes and consequences. *Journal of Clinical Epidemiology* 53:207–16.

Titman, S., K. C. Wei, and F. Xie. 2004. Capital investments and stock returns. *Journal of Financial and Quantitative Analysis* 39:677–700.

Troendle, J. F. 2000. Stepwise normal theory multiple test procedures controlling the false discovery rate. *Journal of Statistical Planning and Inference* 84:139–58.

Todorov, V., and T. Bollerslev. 2010. Jumps and betas: A new framework for disentangling and estimating systematic risks. *Journal of Econometrics* 157:220–35.

Tukey, J. W. 1951. Reminder sheets for “Discussion of paper on multiple comparisons by Henry Scheffe.” In *The Collected Works of John W. Tukey VIII. Multiple Comparisons: 1948-1983*, 469–475. New York: Chapman and Hall.

———. 1953. The problem of multiple comparisons. Unpublished manuscript. In *The Collected Works of John W. Tukey VIII. Multiple Comparisons: 1948-1983*, 1–300. Chapman and Hall, New York.

Tuzel, S. 2010. Corporate real estate holdings and the cross-section of stock returns. *Review of Financial Studies* 23:2268–302.

Valta, P. 2013. Strategic default, debt structure, and stock returns. Working Paper.

van Binsbergen, J. H. 2012. Good-specific habit formation and the cross-section of expected returns. Working Paper.

Vanden, J. M. 2004. Options trading and the CAPM. *Review of Financial Studies* 17:207–38.

———. 2006. Option coskewness and capital asset pricing. *Review of Financial Studies* 19:1279–320.

Vassalou, M. 2003. News related to future GDP growth as a risk factor in equity returns. *Journal of Financial Economics* 68:47–73.

Vassalou, M., and Y. Xing. 2004. Default risk in equity returns. *Journal of Finance* 59:831–68.

Viale, A. M., L. Garcia-Feijoo, and A. Giannetti. 2012. Safety first, robust dynamic asset pricing, and the cross-section of expected stock returns. Working Paper.

Wagenmakers, E. J., and P. Grünwald. 2006. A Bayesian perspective on hypothesis testing: A comment on Killeen (2005). *Psychological Science* 17:641–42.

Wahlen, J. M., and M. M. Wieland. 2011. Can financial statement analysis beat consensus analysts’ recommendations? *Review of Accounting Studies* 16:89–115.

Wang, Yuan. 2012. Debt covenants and cross-sectional equity returns. Working Paper.

Watkins, B. 2003. Riding the wave of sentiment: An analysis of return consistency as a predictor of future returns. *Journal of Behavioral Finance* 4:191–200.

Wei, Z., W. Sun, K. Wang, and H. Hakonarson. 2009. Multiple testing in genome-wide association studies via hidden Markov models. *Bioinformatics* 25:2802–8.

Westfall, P. H., and S. S. Young. 1993. *Resampling-based multiple testing.* New York: John Wiley & Sons.

Welch, I., and A. Goyal. 2008. A comprehensive look at the empirical performance of equity premium prediction. *Review of Financial Studies* 21:1455–508.

White, H. 2000. A reality check for data snooping. *Econometrica* 68:1097–126.

Whited, T. M., and G. Wu. 2006. Financial constraints risk. *Review of Financial Studies* 19:531–59.

Whittemore, A. S. 2007. A Bayesian false discovery rate for multiple testing. *Journal of Applied Statistics* 34:1–9.

Womack, K. L. 1996. Do brokerage analysts' recommendations have investment value? *Journal of Finance* 51:137–67.

Xing, Y. 2008. Interpreting the value effect through the Q-theory: An empirical investigation. *Review of Financial Studies* 21:1767–95.

Xing, Y., X. Zhang, and R. Zhao. 2010. What does the individual option volatility smirk tell us about future equity returns? *Journal of Financial & Quantitative Analysis* 45:641–62.

Yan, S. 2011. Jump risk, stock returns, and slope of implied volatility smile. *Journal of Financial Economics* 99:216–33.

Yekutieli, D., and Y. Benjamini. 1999. Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *Journal of Statistical Planning and Inference* 82:171–96.

Yogo, M. 2006. A consumption-based explanation of expected stock returns. *Journal of Finance* 61:539–80.

Zehetmayer, S., and M. Posch. 2010. Post hoc power estimation in large-scale multiple testing problems. *Bioinformatics* 26:1050–56.

Zhao, X. 2012. Information intensity and the cross-section of stock returns. Working Paper.