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...
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 $R^2$s 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 [1995, 2001] and White [2000]. For research on data snooping and variable selection in predictive regressions,
see Foster, Smith, and Whelely [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.duke.
duke.edu/~charvey/Factor-List.xlsx.

² Other recent papers that systematically study the cross-sectional return patterns include those by Subrahmanyam
[2005] and Green, Hand, and Zhang [2013, 2014]. Other papers that study anomaly discoveries and investor
actions include those by Delarte, Jaffe, and Parmigiani [2014] and Lee et al. [2017].

³ See Jeffreys 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 (1973), 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 Kaplan and Warner (1994) 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), Ahn, Hwang, and Trombley (2000), 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.

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

<table>
<thead>
<tr>
<th>Risk type</th>
<th>Description</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Common</strong> (113)</td>
<td>Proxy for aggregate financial market movement, including market portfolio returns, volatility, squared market returns, among others</td>
<td>Sharpe (1964): market returns; Kwon and Nather (1997): squared market returns</td>
</tr>
<tr>
<td><strong>Financial</strong> (46)</td>
<td>Proxy for movement in macroeconomic fundamentals, including consumption, investment, inflation, among others</td>
<td>Breeden (1979): consumption growth; Cochrane (1999): investment returns</td>
</tr>
<tr>
<td><strong>Macro</strong> (40)</td>
<td>Proxy for aggregate movements in market microstructure or financial market frictions, including liquidity, transaction costs, among others</td>
<td>Pastor and Stambaugh (2003): market liquidity; Lo and Wang (2001): market trading volume</td>
</tr>
<tr>
<td><strong>Accounting</strong> (8)</td>
<td>Proxy for aggregate movements that do not fall into the above categories, including momentum, investors’ beliefs, among others</td>
<td>Pastor and Stambaugh (2003): return momentum; Ang et al. (2006): idiosyncratic volatility</td>
</tr>
<tr>
<td><strong>Financial</strong> (61)</td>
<td>Proxy for firm-level behavioral biases, including analyst dispersion, media coverage, among others</td>
<td>Baker, Miller, and Opler (2003): analyst dispersion; Fama and French (1992): media coverage</td>
</tr>
<tr>
<td><strong>Microstructure</strong> (28)</td>
<td>Proxy for firm-level accounting variables, including PE ratio, debt-to-equity ratio, among others</td>
<td>Basu (1977): PE ratio; Basu (1977): debt-to-equity ratio</td>
</tr>
<tr>
<td><strong>Behavioral</strong> (3)</td>
<td>Proxy for firm-level variables that do not fall into the above categories, including political campaign contributions, ranking-related firm intangibles, among others</td>
<td>Cooper, Gaver, and Ovtchinnikov (2010): political campaign contributions; Cooper, Gaver, and Ovtchinnikov (2010): intangibles</td>
</tr>
</tbody>
</table>

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). 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. 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 Kouwen et al. (2013).
Table 2
Contingency table in testing M hypotheses

Panel A: An example

<table>
<thead>
<tr>
<th></th>
<th>Unpublished</th>
<th>Published</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Truly insignificant</td>
<td>500</td>
<td>50</td>
<td>550</td>
</tr>
<tr>
<td>Truly significant</td>
<td>100</td>
<td>50</td>
<td>150</td>
</tr>
<tr>
<td>Total</td>
<td>600</td>
<td>100(R)</td>
<td>700(M)</td>
</tr>
</tbody>
</table>

Panel B: The testing framework

<table>
<thead>
<tr>
<th></th>
<th>$H_0$ not rejected</th>
<th>$H_0$ rejected</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_0$ true</td>
<td>$N_{0\alpha}$</td>
<td>$N_{0\beta}$</td>
<td>$M_0$</td>
</tr>
<tr>
<td>$H_0$ false</td>
<td>$N_{1\alpha}$</td>
<td>$N_{1\beta}$</td>
<td>$M_1$</td>
</tr>
<tr>
<td>Total</td>
<td>$M - R$</td>
<td>$R$</td>
<td>$M$</td>
</tr>
</tbody>
</table>

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. 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.

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, 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. (2003)].
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. 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, $N_{0r}$ measures the number of rejections when the null is true (i.e., the number of false discoveries) and $N_{1a}$ 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 $\alpha$ is used to control type I error rate: the probability of finding a factor to be significant when it is not. The $\alpha$ is sometimes called the “level of significance.” In a multiple testing framework, restricting each individual test’s type I error rate at $\alpha$ 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 $\alpha$ 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. $N_{0r}$ and $N_{1a}$ count the total number of the two types of errors: $N_{0r}$ counts false discoveries, while $N_{1a}$ 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 (1991). 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.
\( N_{0|r} \). If \( N_{0|r} \) 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 \( \alpha \), we explore two existing methods (Bonferroni and Holm’s adjustment) to ensure FWER does not exceed \( \alpha \). 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:

\[
\text{FDP} = \begin{cases} 
\frac{N_{0|r}}{R} & \text{if } R > 0, \\
0 & \text{if } R = 0.
\end{cases}
\]

The false discovery rate (FDR) is defined as

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

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. Intuitively, this is because FDR allows $N_0 \mid r$ 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 $N_0 = 50 \geq 1$ 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. 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 $N_1 \mid a$ and the fraction of missed discoveries among all abandoned tests $N_1 \mid M-R$ are frequently used to measure the severity of type II errors. 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,

$$FDR = \frac{N_0}{R} \Pr(R > 0) \Pr(R > 0)$$

$$\leq \Pr(N_0 \geq 1) \Pr(R > 0)$$

$$= \Pr(N_0 \geq 1) \cap (R > 0)$$

$$\leq Pr(N_0 \geq 1) \leq FWER,$$

where $I(N_0 \geq 1)$ is an indicator function of event $N_0 \geq 1$. 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 $\alpha$ also controls FDR at $\alpha$.

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, $N(0,1)$; 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, $N(2,1)$. 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 Jager (2008) 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 $\alpha$ 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 $\alpha$.

When comparing two testing procedures that can both achieve a significance level $\alpha$, 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 $\alpha$, we can probably push this error rate closer to $\alpha$ 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.
... and the Cross-Section of Expected Returns

Table 3
A summary of p-value adjustments

<table>
<thead>
<tr>
<th>Adjustment type</th>
<th>Single/Sequential</th>
<th>Multiple test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bonferroni</td>
<td>single</td>
<td>family-wise error rate</td>
</tr>
<tr>
<td>Holm</td>
<td>sequential</td>
<td>family-wise error rate</td>
</tr>
<tr>
<td>Benjamini, Hochberg, and Yekutieli (BHY)</td>
<td>sequential</td>
<td>false discovery rate</td>
</tr>
</tbody>
</table>

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. 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 \( \alpha_w \) and FDR at \( \alpha_d \). 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 \( \alpha_w \) and \( \alpha_d \) at 5%. Table 3 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.

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

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

  \[
p_{i, \text{Bonferroni}} = \min[M \times p_i, 1].
\]

---

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


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

<table>
<thead>
<tr>
<th>Test</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th># of discoveries</th>
</tr>
</thead>
<tbody>
<tr>
<td>t-statistic</td>
<td>1.99</td>
<td>2.63</td>
<td>2.21</td>
<td>3.43</td>
<td>2.17</td>
<td>2.64</td>
<td>4.56</td>
<td>5.34</td>
<td>2.75</td>
<td>2.49</td>
<td>10</td>
</tr>
<tr>
<td>p-value (%)</td>
<td>4.66</td>
<td>0.85</td>
<td>0.05</td>
<td>0.00</td>
<td>0.84</td>
<td>0.00</td>
<td>0.00</td>
<td>0.60</td>
<td>1.28</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Panel B: Bonferroni “significant” factors

<table>
<thead>
<tr>
<th>Test</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>t-statistic</td>
<td>1.99</td>
<td>2.63</td>
<td>2.21</td>
<td>3.43</td>
<td>2.17</td>
<td>2.64</td>
<td>4.56</td>
<td>5.34</td>
<td>2.75</td>
<td>2.49</td>
</tr>
<tr>
<td>p-value (%)</td>
<td>4.66</td>
<td>0.85</td>
<td>0.05</td>
<td>0.00</td>
<td>0.84</td>
<td>0.00</td>
<td>0.00</td>
<td>0.60</td>
<td>1.28</td>
<td></td>
</tr>
</tbody>
</table>

#### Panel C: Holm adjusted p-values and “significant” factors

<table>
<thead>
<tr>
<th>Reordered tests b</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Old order</td>
<td>8</td>
<td>7</td>
<td>4</td>
<td>9</td>
<td>6</td>
<td>2</td>
<td>10</td>
<td>3</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>p-value (%) &amp; # of discoveries</td>
<td>0.00</td>
<td>0.00</td>
<td>0.05</td>
<td>0.60</td>
<td>0.84</td>
<td>0.85</td>
<td>1.28</td>
<td>2.71</td>
<td>3.00</td>
<td>4.66</td>
</tr>
<tr>
<td>αw/(M + 1 - b) &amp; αw = 5%</td>
<td>0.50</td>
<td>0.56</td>
<td>0.63</td>
<td>0.71</td>
<td>0.83</td>
<td>1.00</td>
<td>1.25</td>
<td>1.67</td>
<td>2.50</td>
<td>5.00</td>
</tr>
</tbody>
</table>

#### Panel D: BHY adjusted p-values and “significant” factors

<table>
<thead>
<tr>
<th>Reordered tests b</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Old order</td>
<td>8</td>
<td>7</td>
<td>4</td>
<td>9</td>
<td>6</td>
<td>2</td>
<td>10</td>
<td>3</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>p-value (%) &amp; # of discoveries</td>
<td>0.00</td>
<td>0.00</td>
<td>0.05</td>
<td>0.60</td>
<td>0.84</td>
<td>0.85</td>
<td>1.28</td>
<td>2.71</td>
<td>3.00</td>
<td>4.66</td>
</tr>
<tr>
<td>(b·αd)/(M × c(M)) &amp; αd = 5%</td>
<td>0.17</td>
<td>0.34</td>
<td>0.51</td>
<td>0.68</td>
<td>0.85</td>
<td>1.02</td>
<td>1.19</td>
<td>1.37</td>
<td>1.54</td>
<td>1.71</td>
</tr>
</tbody>
</table>

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) = ∑Mj=1 1/j. b is the order of p-values from lowest to highest. αw is the significance level for Bonferroni’s and Holm’s procedure, and αd 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 αw.

**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 αw = 5%. Equivalently, we can look at the original p-values and consider the cutoff of 0.5%(= αw/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 4 and assuming M0 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 (M0 × αw)/M, without any assumption on the dependence structure of the p-values. Since M0 ≤ M, Bonferroni also controls FWER at level αw.

---

21 The number of true nulls M0 is unknown, so we usually cannot make Bonferroni more powerful by increasing αw to α = Mαw/M0 (note that Mαw/M = αw). 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. They are motivated by a seminal paper by Schweder and Spjøtvoll (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 Spjøtvoll (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 \( p_1 \leq p_2 \leq \cdots \leq p_M \), and let the associated null hypotheses be \( H_1, H_2, \cdots, H_M \).
- Let \( k \) be the minimum index such that \( p_k > \frac{\alpha_0}{M + 1 - b} \).
- Reject the null hypotheses \( H_1, H_2, \cdots, H_{k-1} \) (i.e., declare these factors significant), but not \( H_k, \cdots, H_M \).

The equivalent adjusted \( p \)-value is therefore

\[
p_{(j)}^{Holm} = \min\{\max\{(M - j + 1)p_{(j)}\}, 1\}.
\]

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. If \( k \) is the smallest index that satisfies \( p_k > \frac{\alpha_0}{M + 1 - b} \), 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 \( p_k > \frac{\alpha_0}{M + 1 - b} \). This means that for \( b < k \), \( p_b \leq \frac{\alpha_0}{M + 1 - b} \). In particular, for \( b = 1 \), Bonferroni equals Holm, that is, \( \frac{\alpha_0}{M} = \frac{\alpha_0}{M + 1 - b} \) for \( b = 2 \).

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

\[\text{Here, “sequential” refers to the fact that we adjust the ordered } p \text{-values sequentially. It does not mean that the individual tests arrive sequentially.}
\]

\[\text{Viewing small } p \text{-values as “up” and large } p \text{-values as “down,” Holm’s procedure is a “step-down” procedure in that it goes from small } p \text{-values to large ones. This terminology is consistent with the statistics literature. Of course, small } p \text{-values are associated with “large” values of the test statistics.}
\]
α_M < α_{M+1−b} \Rightarrow \text{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 3, we first order the p-values in ascending order and try to locate the smallest index k that makes \( p(b) > \frac{α_M}{M+1−b} \). Table 3 panel C, shows the ordered p-values and the associated \( \frac{α_M}{M+1−b} \)'s. Starting from the smallest p-value and going up, we see that \( p(5) \) is below \( \frac{α_M}{M+1−5} \) until \( b=5 \), at which point \( p(5) \) is above \( \frac{α_M}{10+1−5} \). Therefore, the smallest \( b \) that satisfies \( p(b) > \frac{α_M}{M+1−b} \) 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 \( p(1) \leq p(2) \leq \cdots \leq p(b) \leq \cdots \leq p(M) \) and let associated null hypotheses be \( H(1), H(2), \cdots, H(b), \cdots, H(M) \).
- Let \( k \) be the maximum index such that \( p(b) \leq \frac{b}{M \times c(M)} α_d \).
- 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

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

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

\[
c(M) = \sum_{j=1}^{M} \frac{1}{j}
\]

(1)
... and the Cross-Section of Expected Returns

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 \( \alpha_d \) (significance level for FDR) is chosen to be the same as \( \alpha_w \) (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 \( p(k+1) \) is below \( \frac{k+1}{M \times c(M)} \alpha_d \). This means that for \( b > k \), \( p(b) > \frac{b}{M \times c(M)} \alpha_d \). In particular, we have

\[
p(k+1) > \frac{k+1}{M \times c(M)} \alpha_d, \quad p(k+2) > \frac{k+2}{M \times c(M)} \alpha_d, \quad \ldots, \quad p(M) > \frac{M}{M \times c(M)} \alpha_d.
\]

We see that the \((k+1)\)th to the last null hypotheses, not rejected, are compared to numbers smaller than \( \alpha_d \), 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 \( p(k) \leq \frac{k}{M \times c(M)} \alpha_d \). Table 4, panel D, shows the ordered \( p \)-values and the associated \( \frac{k}{M \times c(M)} \alpha_d \)'s. Starting from the largest \( p \)-value and going down, we see that \( p(6) \) is above \( \frac{6}{10 \times c(M)} \alpha_d \) until \( b = 6 \), at which point \( p(6) \) is below \( \frac{6}{10 \times c(M)} \alpha_d \). 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
Ordered tests (b)

p−value

Bonferroni (3 discoveries)

Single tests
(10 discoveries)

Holm

BHY
(6 discoveries)

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. Note with c(M) > 1, this reduces the size of \( \frac{b}{(M+c(M))n} \alpha_d \) and it is tougher to satisfy the inequality \( p(b) \leq \frac{b}{(M+c(M))n} \alpha_d \). 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. 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 ‡.
… and the Cross-Section of Expected Returns

![Figure 2](image)

**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. The overwhelming majority of t-statistics exceed the 1.96 benchmark for 5% significance. 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 \(M = R\) 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. 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, 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 \(\alpha_w\) at 5% (Holm, FWER) and \(\alpha_d\) 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 \(\alpha_w\) at 5%. On the other hand, FDR is a more lenient control relative to FWER. If we choose the same \(\alpha_d\) as \(\alpha_w\), 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 \(\alpha_d\) 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
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 French 1992), 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 2006), and CVOL (consumption volatility; Breeden and Grundy 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.xls.

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 French 1992), 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 2006), and CVOL (consumption volatility; Breeden and Grundy 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.xls.
(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 \( \alpha \), the law of large numbers forces the false discovery rate (FDR) to converge to a constant. If we change \( \alpha_d \) 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. 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.


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 $\alpha$ 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 I). 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.

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.
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 $\gamma$ 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. 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.

---

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:

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

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 $\mu_i$ denotes the population mean of the strategy. The $\mu_i$’s are unobservable, and hypothesis testing under this framework amounts to testing $\mu_i > 0$. We assume that each $\mu_i$ 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, $\text{Exp}(\lambda)$ is the exponential distribution that has a mean parameter $\lambda$, 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 Kim (2004), Kosowsky et al. (2006), Kosowsky, Nair, and Lee (2009), Lam and Frankel (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 (2014b) 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 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.

As a starting point, we assume that the contemporaneous correlation between two strategies’ returns is \( \rho \). The noncontemporaneous correlations are assumed to be zero. That is,

\[
\text{Corr}(X_{i,t}, X_{j,t}) = \rho, \quad i \neq j,
\]

\[
\text{Corr}(X_{i,t}, X_{j,s}) = 0, \quad t \neq s.
\]

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 sample.

---

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:

\[
\begin{align*}
\hat{T} & = \text{Total number of discoveries} = 353, \\
\hat{Q}_1 & = \text{The 20th percentile of the sample of } t\text{-statistics} = 2.39, \\
\hat{Q}_2 & = \text{The 50th percentile of the sample of } t\text{-statistics} = 3.16, \\
\hat{Q}_3 & = \text{The 90th percentile of the sample of } t\text{-statistics} = 6.34.
\end{align*}
\]

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:

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

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, p0, 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, p0, 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, p0, 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 $\mu_i$ 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 [Eisen 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)

<table>
<thead>
<tr>
<th>$\rho$</th>
<th>$p_0$</th>
<th>$\lambda$</th>
<th>$M$</th>
<th>FWER(5%)</th>
<th>FWER(1%)</th>
<th>FDR(5%)</th>
<th>FDR(1%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.396</td>
<td>0.550</td>
<td>1,297</td>
<td>3.89</td>
<td>4.28</td>
<td>2.16</td>
<td>2.88</td>
</tr>
<tr>
<td>0.2</td>
<td>0.444</td>
<td>0.555</td>
<td>1,378</td>
<td>3.91</td>
<td>4.30</td>
<td>2.57</td>
<td>2.95</td>
</tr>
<tr>
<td>0.4</td>
<td>0.485</td>
<td>0.554</td>
<td>1,477</td>
<td>3.81</td>
<td>4.23</td>
<td>2.34</td>
<td>3.05</td>
</tr>
<tr>
<td>0.6</td>
<td>0.601</td>
<td>0.555</td>
<td>1,775</td>
<td>3.67</td>
<td>4.15</td>
<td>2.43</td>
<td>3.09</td>
</tr>
<tr>
<td>0.8</td>
<td>0.840</td>
<td>0.560</td>
<td>3,110</td>
<td>3.35</td>
<td>3.89</td>
<td>2.59</td>
<td>3.25</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>$\rho$</th>
<th>$p_0$</th>
<th>$\lambda$</th>
<th>$M$</th>
<th>FWER(5%)</th>
<th>FWER(1%)</th>
<th>FDR(5%)</th>
<th>FDR(1%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.683</td>
<td>0.550</td>
<td>2,458</td>
<td>4.17</td>
<td>4.55</td>
<td>2.69</td>
<td>3.30</td>
</tr>
<tr>
<td>0.2</td>
<td>0.722</td>
<td>0.551</td>
<td>2,696</td>
<td>4.15</td>
<td>4.54</td>
<td>2.76</td>
<td>3.38</td>
</tr>
<tr>
<td>0.4</td>
<td>0.773</td>
<td>0.552</td>
<td>3,031</td>
<td>4.06</td>
<td>4.45</td>
<td>2.80</td>
<td>3.40</td>
</tr>
<tr>
<td>0.6</td>
<td>0.885</td>
<td>0.562</td>
<td>4,339</td>
<td>3.86</td>
<td>4.29</td>
<td>2.91</td>
<td>3.55</td>
</tr>
<tr>
<td>0.8</td>
<td>0.922</td>
<td>0.532</td>
<td>5,392</td>
<td>3.44</td>
<td>4.00</td>
<td>2.75</td>
<td>3.39</td>
</tr>
</tbody>
</table>

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 $r=2/3$. $p_0$ is the proportion of having a strategy that has a mean of zero. $\rho$ 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).

For the other parameter estimates, both $p_0$ and $M$ are increasing in $\rho$. 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 $\rho$ 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,

---

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. 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 $\rho$?

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 $\rho$.

First, the value of the optimized objective function sheds light on the level of $\rho$. Intuitively, a value of $\rho$ that is more consistent with the data-generating process should result in a lower optimized objective function. Across the various specifications of $\rho$ 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 $\rho$ 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 (2011) 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.

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 Bonferroni, 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

<table>
<thead>
<tr>
<th>Reference</th>
<th>Factor</th>
<th>#</th>
<th>Reference</th>
<th>Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sharpe (1964)</td>
<td>market return</td>
<td>T</td>
<td>Constantinides (1982)</td>
<td>individual consumer’s wealth</td>
</tr>
<tr>
<td>Fama (1965)</td>
<td>market return</td>
<td>T</td>
<td>Jensen (1966)</td>
<td>EP ratio</td>
</tr>
<tr>
<td>Constantinides (1967)</td>
<td>total volatility</td>
<td>C1</td>
<td>Black, Jensen, and Scholes (1972)</td>
<td>institutional holding†</td>
</tr>
<tr>
<td>Hecker (1972)</td>
<td>market return</td>
<td>T</td>
<td>Black (1973)</td>
<td>relative prices of cons. goods</td>
</tr>
<tr>
<td>Black (1973)</td>
<td>market return</td>
<td>T</td>
<td>Black (1973)</td>
<td>institutional holding</td>
</tr>
<tr>
<td>Fama and French (1973)</td>
<td>state variables investment opps.</td>
<td>T</td>
<td>De Bondt and Thaler (1985)</td>
<td>change in expected inflation*</td>
</tr>
<tr>
<td>Sharpe (1972)</td>
<td>high-order market return</td>
<td>T</td>
<td>De Bondt and Thaler (1985)</td>
<td>unanticipated inflation</td>
</tr>
<tr>
<td>Reinganum (1974)</td>
<td>individual investor resources</td>
<td>T</td>
<td>De Bondt and Thaler (1985)</td>
<td>term structure†</td>
</tr>
<tr>
<td>Garbade and Ofer (1975)</td>
<td>earnings growth expectations</td>
<td>C3</td>
<td>De Bondt and Thaler (1985)</td>
<td>transaction costs</td>
</tr>
<tr>
<td>Kraus and Litzenberger (1976)</td>
<td>market return†</td>
<td>T</td>
<td>De Bondt and Thaler (1985)</td>
<td>transaction costs</td>
</tr>
<tr>
<td>Berk (1977)</td>
<td>PE ratio</td>
<td>C4</td>
<td>De Bondt and Thaler (1985)</td>
<td>long-term interest rate</td>
</tr>
<tr>
<td>Lucas (1978)</td>
<td>marginal rate of substitution</td>
<td>T</td>
<td>De Bondt and Thaler (1985)</td>
<td>industrial production growth†</td>
</tr>
<tr>
<td>Litterman and Ross (1979)</td>
<td>dividend yield</td>
<td>C5</td>
<td>De Bondt and Thaler (1985)</td>
<td>credit premium†</td>
</tr>
<tr>
<td>Brennan (1979)</td>
<td>real consumption growth</td>
<td>T</td>
<td>De Bondt and Thaler (1985)</td>
<td>term structure†</td>
</tr>
<tr>
<td>Fama (1981)</td>
<td>short interest</td>
<td>C7</td>
<td>(continued)</td>
<td></td>
</tr>
</tbody>
</table>
Table 6

<table>
<thead>
<tr>
<th>Reference</th>
<th>Factor</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ferson and Harvey (1991)</td>
<td>market return †</td>
<td>change in expected inflation</td>
</tr>
<tr>
<td>Ferson and Harvey (1991)</td>
<td>consumption growth †</td>
<td>change in expected GNP</td>
</tr>
<tr>
<td>Ferson and Harvey (1991)</td>
<td>credit spread †</td>
<td>change in expected equity offerings</td>
</tr>
<tr>
<td>Ferson and Harvey (1991)</td>
<td>value-added inflation</td>
<td>change in expected inflation</td>
</tr>
<tr>
<td>Ferson and Harvey (1991)</td>
<td>real short rate</td>
<td>change in expected inflation</td>
</tr>
<tr>
<td>Ferson and Harvey (1991)</td>
<td>term structure</td>
<td>change in expected inflation</td>
</tr>
<tr>
<td>Ferson and Harvey (1991)</td>
<td>returns on S&amp;P stocks</td>
<td>change in expected inflation</td>
</tr>
<tr>
<td>Ferson and Harvey (1991)</td>
<td>value</td>
<td>change in expected inflation</td>
</tr>
<tr>
<td>Ferson and Harvey (1991)</td>
<td>size</td>
<td>change in expected inflation</td>
</tr>
<tr>
<td>Ferson and Harvey (1991)</td>
<td>dividend yield †</td>
<td>change in expected inflation</td>
</tr>
<tr>
<td>Chopra, Lakonishok, and Ritter (1992)</td>
<td>return momentum ‡</td>
<td>change in expected inflation</td>
</tr>
<tr>
<td>Holthausen and Larcker (1992)</td>
<td>predicted return signs ‡</td>
<td>change in expected inflation</td>
</tr>
<tr>
<td>Jegadeesh and Titman (1993)</td>
<td>return momentum</td>
<td>change in expected inflation</td>
</tr>
<tr>
<td>Elton et al. (1993)</td>
<td>returns on S&amp;P stocks ‡</td>
<td>change in expected inflation</td>
</tr>
<tr>
<td>Elton et al. (1993)</td>
<td>returns on non-S&amp;P stocks ‡</td>
<td>change in expected inflation</td>
</tr>
<tr>
<td>Bansal and Viswanathan (1993)</td>
<td>high-order equity &amp; bond returns ‡</td>
<td>change in expected inflation</td>
</tr>
<tr>
<td>La Porta (1996)</td>
<td>earnings forecasts</td>
<td>change in expected inflation</td>
</tr>
<tr>
<td>Fama and French (1993)</td>
<td>market return †</td>
<td>change in expected inflation</td>
</tr>
<tr>
<td>Lev and Sougiannis (1996)</td>
<td>R&amp;D capital</td>
<td>change in expected inflation</td>
</tr>
<tr>
<td>Sloan (1996)</td>
<td>accruals</td>
<td>change in expected inflation</td>
</tr>
<tr>
<td>Womack (1996)</td>
<td>buy recommendations</td>
<td>change in expected inflation</td>
</tr>
<tr>
<td></td>
<td>sell recommendations</td>
<td>change in expected inflation</td>
</tr>
<tr>
<td></td>
<td>credit risk †</td>
<td>change in expected inflation</td>
</tr>
<tr>
<td>Erb, Harvey, and Viskanta (1996)</td>
<td>credit rating</td>
<td>change in expected inflation</td>
</tr>
<tr>
<td>Schiller and Stambaugh (1996)</td>
<td>LT inflation expectations ‡</td>
<td>change in expected inflation</td>
</tr>
<tr>
<td>Chapman (1997)</td>
<td>nonlinear fn. of cons. growth ‡</td>
<td>change in expected inflation</td>
</tr>
<tr>
<td>Fung and Hsieh (1997)</td>
<td>opportunistic style return ‡</td>
<td>change in expected inflation</td>
</tr>
<tr>
<td>Fung and Hsieh (1997)</td>
<td>global/macro style return ‡</td>
<td>change in expected inflation</td>
</tr>
<tr>
<td>Chang (1997)</td>
<td>distressed inv. style return ‡</td>
<td>change in expected inflation</td>
</tr>
<tr>
<td>Daniel and Titman (1997)</td>
<td>size †</td>
<td>change in expected inflation</td>
</tr>
<tr>
<td>Loughran and Ritter (1995)</td>
<td>new public stock issuance</td>
<td>change in expected inflation</td>
</tr>
<tr>
<td>Michaely, Thaler, and Womack (1995)</td>
<td>dividend initiations</td>
<td>change in expected inflation</td>
</tr>
<tr>
<td></td>
<td>dividend omissions</td>
<td>change in expected inflation</td>
</tr>
<tr>
<td>Ackert and Athanassakos (1997)</td>
<td>earnings forecast uncertainty</td>
<td>change in expected inflation</td>
</tr>
<tr>
<td>Botosan (1997)</td>
<td>disclosure level</td>
<td>change in expected inflation</td>
</tr>
<tr>
<td>Bossaerts and Dammon (1994)</td>
<td>tax rate for capital gains</td>
<td>change in expected inflation</td>
</tr>
<tr>
<td>Ackert and Athanassakos (1997)</td>
<td>tax rate for dividends</td>
<td>change in expected inflation</td>
</tr>
</tbody>
</table>

(continued)
<table>
<thead>
<tr>
<th>Reference</th>
<th>Factor</th>
<th>#</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beneish (1997)</td>
<td>earnings management likelihood$^\text{a}$</td>
<td>C27</td>
</tr>
<tr>
<td>Loughran and Vijh (1997)</td>
<td>corporate acquisitions</td>
<td></td>
</tr>
<tr>
<td>Brennan, Chordia, and Subrahmanyam (1998)</td>
<td>book-to-market ratio$^\text{a}$</td>
<td></td>
</tr>
<tr>
<td>Abarbanell and Bushee (1998)</td>
<td>trading volume</td>
<td>C28</td>
</tr>
<tr>
<td>Bakker and Lee (1998)</td>
<td>fundamental analysis$^\text{a}$</td>
<td></td>
</tr>
<tr>
<td>Diether, Malloy, and Scherbina (2002)</td>
<td>bankruptcy risk</td>
<td>C29</td>
</tr>
<tr>
<td>Diether (1998)</td>
<td>illiquidity</td>
<td>C30</td>
</tr>
<tr>
<td>Datar, Naik, and Radcliffe (1998)</td>
<td>expected portfolio return</td>
<td>F28</td>
</tr>
<tr>
<td>Ferson and Harvey (1999)</td>
<td>industry momentum</td>
<td>C31</td>
</tr>
<tr>
<td>Moskowitz and Grinblatt (1999)</td>
<td>debt offerings$^\text{a}$</td>
<td>F29</td>
</tr>
<tr>
<td>Springs and Affleck-Graves (1999)</td>
<td>entrepreneur income</td>
<td>C32</td>
</tr>
<tr>
<td>Horvath and Loewy (2000)</td>
<td>cost of equity</td>
<td>C33</td>
</tr>
<tr>
<td>Harvey and Siddique (2000)</td>
<td>trading volume</td>
<td>C34</td>
</tr>
<tr>
<td>Lee and Swaminathan (2000)</td>
<td>intra-industry value</td>
<td>C35</td>
</tr>
<tr>
<td>Asness, Pover, and Verrecchia (2000)</td>
<td>intra-industry C/P</td>
<td>C36</td>
</tr>
<tr>
<td>Poterba (2000)</td>
<td>intra-industry momentum</td>
<td>C37</td>
</tr>
<tr>
<td>Lettau and Ludvigson (2001)</td>
<td>financial statement information$^\text{a}$</td>
<td>C38</td>
</tr>
<tr>
<td>Claessens, Subrahmanyam, and Ambachtsheer (2001)</td>
<td>consumption growth$^\text{a}$</td>
<td></td>
</tr>
<tr>
<td>Largent, Bub, and S-R (2001)</td>
<td>consumption-wealth ratio</td>
<td>F31</td>
</tr>
<tr>
<td>Fung and Hsieh (2001)</td>
<td>level of liquidity</td>
<td>C39</td>
</tr>
<tr>
<td>Barber et al. (2001)</td>
<td>variability of liquidity</td>
<td>C40</td>
</tr>
<tr>
<td>Dichev (2001)</td>
<td>financial constraints</td>
<td>C41</td>
</tr>
<tr>
<td>Fama and French (2001)</td>
<td>straddle return$^\text{a}$</td>
<td>C42</td>
</tr>
<tr>
<td>Diether and Poterba (2001)</td>
<td>bond rating changes</td>
<td>C43</td>
</tr>
<tr>
<td>Fama and French (2001)</td>
<td>consensus recommendations</td>
<td>C44</td>
</tr>
<tr>
<td>Goetzmann and Metrick (2001)</td>
<td>institutional ownership</td>
<td></td>
</tr>
<tr>
<td>Dittmar (2002)</td>
<td>market return</td>
<td>C45</td>
</tr>
<tr>
<td>Poterba (2000)</td>
<td>squared market return</td>
<td>C46</td>
</tr>
<tr>
<td>Lettau and Ludvigson (2001)</td>
<td>labor income growth$^\text{a}$</td>
<td></td>
</tr>
<tr>
<td>Claessens, Subrahmanyam, and Ambachtsheer (2001)</td>
<td>squared labor income growth</td>
<td>F32</td>
</tr>
</tbody>
</table>

(continued)
<table>
<thead>
<tr>
<th>Factor</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>F46</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>C60</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>F47</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>C64</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>F62</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>C65</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>F66</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>C67</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>F68</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>C68</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>F69</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>C69</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>F70</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>C70</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>F71</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>C71</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>F72</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>C72</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>F73</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>C73</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>F74</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>C74</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>F75</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>C75</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>F76</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>C76</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>F77</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>C77</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>F78</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>C78</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>F79</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>C79</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>F80</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>C80</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>F81</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>C81</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>F82</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>C82</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>F83</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>C83</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>F84</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>C84</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>F85</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>C85</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>F86</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>C86</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>F87</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>C87</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>F88</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>C88</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>F89</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>C89</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>F90</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>C90</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>F91</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>C91</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>F92</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>C92</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>F93</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>C93</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>F94</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>C94</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>F95</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>C95</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>F96</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>C96</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>F97</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>C97</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>F98</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>C98</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>F99</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>C99</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>F100</td>
</tr>
<tr>
<td>Housing, Rental, and Rentable Income (2000)</td>
<td>C100</td>
</tr>
<tr>
<td>Reference Factor</td>
<td>Factor</td>
</tr>
<tr>
<td>------------------</td>
<td>--------</td>
</tr>
<tr>
<td>Cooper, Gulen, and Schill (2008) asset growth</td>
<td>C101</td>
</tr>
<tr>
<td>Barber, Odean, and Zhu (2009) order imbalance</td>
<td>C125</td>
</tr>
<tr>
<td>Pontiff and Woodgate (2008) share issuance</td>
<td>C102</td>
</tr>
<tr>
<td>Cremers, Halling, and Weinbaum (2010) market volatility and jumps</td>
<td>F85</td>
</tr>
<tr>
<td>Brandt et al. (2008) earnings announcement return</td>
<td>C103</td>
</tr>
<tr>
<td>Hirshleifer and Jiang (2010) market mispricing</td>
<td>F86</td>
</tr>
<tr>
<td>Cohen and Frazzini (2008) firm economic links</td>
<td>C104</td>
</tr>
<tr>
<td>Boyer, Mitton, and Vorkink (2010) idiosyncratic skewness</td>
<td>C126</td>
</tr>
<tr>
<td>Fabozzi, ... vol.</td>
<td>C133</td>
</tr>
<tr>
<td>Malloy, ... (2009) LT stockholder cons. growth</td>
<td>C136</td>
</tr>
<tr>
<td>Cremers, Nair, and John (2009) takeover likelihood</td>
<td>F80</td>
</tr>
<tr>
<td>net cash to equity holders</td>
<td>C137</td>
</tr>
<tr>
<td>Chordia, Huh, and Subrahmanyam ... (2012) extreme downside risk</td>
<td>C139</td>
</tr>
<tr>
<td>Kapadia (2011) distress risk</td>
<td>C125</td>
</tr>
<tr>
<td>Fu (2009) idiosyncratic volatility</td>
<td>C112</td>
</tr>
<tr>
<td>Hou, Karolyi, and Kho (2011) momentum</td>
<td>†</td>
</tr>
<tr>
<td>Hahn and Lee (2009) debt capacity</td>
<td>C113</td>
</tr>
<tr>
<td>cash flow-to-price</td>
<td>F89</td>
</tr>
<tr>
<td>Bali and Hovakimian (2009) realized-implied vol. spread</td>
<td>C114</td>
</tr>
<tr>
<td>Li (2011) R&amp;D investment</td>
<td>C141</td>
</tr>
<tr>
<td>Chandrashekar and Rao (2009) productivity of cash</td>
<td>C116</td>
</tr>
<tr>
<td>Bali, Cakici, and Whitelaw (2011) extreme stock returns</td>
<td>C142</td>
</tr>
<tr>
<td>Chemmanur and Yan (2009) advertising</td>
<td>C117</td>
</tr>
<tr>
<td>Yan (2011) jumps individual stock returns</td>
<td>C143</td>
</tr>
<tr>
<td>Da and Warachka (2009) analyst forecasts optimism</td>
<td>C118</td>
</tr>
<tr>
<td>Gokcen (2009) information revelation</td>
<td>C119</td>
</tr>
<tr>
<td>Chen, Novy-Marx, and Zhang (2011) market return</td>
<td>†</td>
</tr>
<tr>
<td>Gow and Taylor (2009) earnings volatility</td>
<td>C120</td>
</tr>
<tr>
<td>Huang (2009) cash-flow volatility</td>
<td>C121</td>
</tr>
<tr>
<td>ROE portfolio return</td>
<td>F90</td>
</tr>
<tr>
<td>Korniotis and Kumar (2009) local unemployment</td>
<td>C122</td>
</tr>
<tr>
<td>Nguyen and Swanson (2009) efficiency score</td>
<td>C124</td>
</tr>
<tr>
<td>Han and Zhou (2011) credit default swap spreads</td>
<td>C147</td>
</tr>
<tr>
<td>(continued)</td>
<td></td>
</tr>
<tr>
<td>Reference</td>
<td>Factor</td>
</tr>
<tr>
<td>-----------</td>
<td>--------</td>
</tr>
<tr>
<td>Eisfeldt and Paramikoian (2011)</td>
<td>organizational capital</td>
</tr>
<tr>
<td>Balakrishnan and Mohammadi (2011)</td>
<td>residual income</td>
</tr>
<tr>
<td>Gharibov and Huang (2010)</td>
<td>accrual volatility</td>
</tr>
<tr>
<td>Cullen and Lynch (2011)</td>
<td>implied cost of capital</td>
</tr>
<tr>
<td>Cullen, Khun, and Li (2011)</td>
<td>nonaccounting info. quality</td>
</tr>
<tr>
<td>Chen, Kasznieriak, and Ortiz-Molina (2011)</td>
<td>accounting info. quality</td>
</tr>
<tr>
<td>Dr. Liu and Schumacker (2011)</td>
<td>labor unions</td>
</tr>
<tr>
<td>Drake, Roes, and Swaminathan (2011)</td>
<td>expected return uncertainty</td>
</tr>
<tr>
<td>Halpern, Rabinovitch, and Van Wier (2011)</td>
<td>short interest</td>
</tr>
<tr>
<td>Hess, Kreigström, and Pächer (2011)</td>
<td>percent total accrual</td>
</tr>
<tr>
<td>Imboden and Tozat (2011)</td>
<td>projected earnings accuracy</td>
</tr>
<tr>
<td>Landisman et al. (2011)</td>
<td>firm productivity</td>
</tr>
<tr>
<td>Li (2011)</td>
<td>earnings forecast</td>
</tr>
<tr>
<td>Nohria and Power (2011)</td>
<td>asset growth</td>
</tr>
<tr>
<td>Ortiz-Molina and Phillips (2011)</td>
<td>real asset liquidity</td>
</tr>
<tr>
<td>Panagopoulos (2011)</td>
<td>customer-base concentration</td>
</tr>
<tr>
<td>Wahlen and Wier (2011)</td>
<td>tax expense surprises</td>
</tr>
<tr>
<td>Garbade and Yip (2011)</td>
<td>shareholder recovery</td>
</tr>
<tr>
<td>Saxon (2011)</td>
<td>consumption increase score</td>
</tr>
<tr>
<td>Abbott, Jena, and Maiz (2012)</td>
<td>garbage growth</td>
</tr>
<tr>
<td>Campbell et al. (2012)</td>
<td>financial intermediary’s wealth</td>
</tr>
<tr>
<td>Chen and Pederson (2013)</td>
<td>stochastic volatility</td>
</tr>
<tr>
<td>Shin (2013)</td>
<td>average variance of equity returns</td>
</tr>
<tr>
<td>Benartzi and Kosiborod (2013)</td>
<td>income growth for manufacturing</td>
</tr>
<tr>
<td>Chua, Davis, and Jansen (2013)</td>
<td>income growth for government</td>
</tr>
<tr>
<td>Wahlen and Wier (2011)</td>
<td>income growth for services</td>
</tr>
<tr>
<td>Balai and Zhou (2013)</td>
<td>income growth for consumption</td>
</tr>
<tr>
<td>Boonen and Lowery-Braun (2013)</td>
<td>income growth for production</td>
</tr>
<tr>
<td>Boonen, Lee, and Nichols (2012)</td>
<td>labor income</td>
</tr>
<tr>
<td>Fang (2011)</td>
<td>income growth for production</td>
</tr>
<tr>
<td>Fang, Lee, and Nichols (2012)</td>
<td>labor income</td>
</tr>
<tr>
<td>Fang and Chui (2014)</td>
<td>income growth for consumption</td>
</tr>
<tr>
<td>Fang, Lee, and Nichols (2012)</td>
<td>consumption growth</td>
</tr>
<tr>
<td>Fang and Chui (2014)</td>
<td>income growth for services</td>
</tr>
<tr>
<td>Fang, Lee, and Nichols (2012)</td>
<td>income growth for government</td>
</tr>
<tr>
<td>Fang and Chui (2014)</td>
<td>earnings growth</td>
</tr>
<tr>
<td>Fang, Lee, and Nichols (2012)</td>
<td>earnings growth</td>
</tr>
<tr>
<td>Fang and Chui (2014)</td>
<td>earnings growth</td>
</tr>
<tr>
<td>Fang, Lee, and Nichols (2012)</td>
<td>earnings growth</td>
</tr>
<tr>
<td>Fang and Chui (2014)</td>
<td>earnings growth</td>
</tr>
<tr>
<td>Fang, Lee, and Nichols (2012)</td>
<td>earnings growth</td>
</tr>
<tr>
<td>Fang, Lee, and Nichols (2012)</td>
<td>earnings growth</td>
</tr>
<tr>
<td>Fang and Chui (2014)</td>
<td>earnings growth</td>
</tr>
<tr>
<td>Fang, Lee, and Nichols (2012)</td>
<td>earnings growth</td>
</tr>
<tr>
<td>Fang and Chui (2014)</td>
<td>earnings growth</td>
</tr>
</tbody>
</table>
Table 6
Continued

<table>
<thead>
<tr>
<th>Reference</th>
<th>Factor</th>
<th>#</th>
<th>Reference</th>
<th>Factor</th>
<th>#</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cohen, Diether, and Malloy (2013)</td>
<td>firm’s ability to innovate</td>
<td>C194</td>
<td>Abbas et al. (2013)</td>
<td>convertible debt indicator</td>
<td>C200</td>
</tr>
</tbody>
</table>

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/~charvey/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.

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 their return factors.

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

8: Maksimovic (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 of Kumbhakar (2003). This is different from the liquidity measure of 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 return factors from the long-run risk model.

12: No \( p \)-values reported for their distress tracking factor.

13: Gómez, Pal, and Zapatero (2013) study census division level labor income. However, most of the division level labor income have a non-significant \( t \)-statistic. We do not count their factor.

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, 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. 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. 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. We choose to use the simplest distribution that incorporates this monotonicity condition: the exponential distribution.

46 See Heckman (1979) and Greene (2008), Chapter 24).
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.
... and the Cross-Section of Expected Returns

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

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

where $\Phi(c; \lambda)$ is the cumulative distribution function evaluated at $c$ for a exponential distribution with mean $\lambda$. 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. 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 $\hat{\lambda}$ and $\hat{M}$ 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 $\hat{\lambda}$ 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

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

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.

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

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

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 \(\lambda\) 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, 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.
... and the Cross-Section of Expected Returns

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 (2002) and highlights the key issues in Bayesian multiple hypothesis testing. More sophisticated generalizations modify the basic model but are unlikely to change the fundamental hierarchical testing structure. We use this framework to explain the pros and cons of performing multiple testing in a Bayesian framework.

The hierarchical model is as follows:

\[
\begin{align*}
H_1: \quad & (X_i | \mu_i, \sigma_i^2, \gamma_i) \sim N(\gamma_i \mu_i, \sigma_i^2), \\
H_2: \quad & \mu_i | \tau_1^2 \sim N(0, \tau_1^2), \gamma_i | p_0 \sim \text{Ber}(1 - p_0), \\
H_3: \quad & (\tau_1^2, \sigma_i^2) \sim \pi(\tau_1^2, \sigma_i^2), p_0 \sim \pi_2(p_0).
\end{align*}
\]

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; \( \mu_i \) is the unknown mean return; \( \sigma_i^2 \) is the common variance for returns; and \( \gamma_i \) is an indicator function, with \( \gamma_i = 0 \) indicating a zero factor mean. \( \gamma_i \) 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 \( \sigma_i^2 \), 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—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 \( \mu_i \)'s and \( \gamma_i \)'s are assumed to be generated from two other parametric distributions: \( \mu_i \)'s are independently generated from a normal distribution, and \( \gamma_i \)'s are simply generated from a Bernoulli distribution, that is, \( \gamma_i = 0 \) with probability \( p_0 \).

The normality assumption for the \( \mu_i \)'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’ method—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. (2004) and Efron (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 (2004).

56 See Meng and Dempster (1992) 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 $\tau^2$ and $\sigma^2$ follow a joint prior distribution $\pi_1$ and the probability $\pi_0$ follows a prior distribution $\pi_2$.

Objective or “neutral” priors for $\pi_1$ and $\pi_2$ can be specified as:

$$\pi_1(\tau^2, \sigma^2) \propto (\tau^2 + \sigma^2)^{-2},$$
$$\pi_2(\pi_0) = \text{Uniform}(0, 1).$$

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” automatically adjusts the posterior probabilities when more factors are simultaneously tested. 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. These technical issues create additional hurdles for the application of the Bayesian approach.

References


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


… and the Cross-Section of Expected Returns


... and the Cross-Section of Expected Returns


… and the Cross-Section of Expected Returns


and the Cross-Section of Expected Returns


… and the Cross-Section of Expected Returns


64
... and the Cross-Section of Expected Returns


… and the Cross-Section of Expected Returns


67


