Journal of Financial Economics 141 (2021) 413–435

Contents lists available at ScienceDirect

Journal of Financial Economics

journal homepage: www.elsevier.com/locate/jfec

Lucky factors[☆]

Campbell R. Harvey[a,b], Yan Liu[c,*]

[a] Duke University, 100 Fuqua Drive, Durham, NC 27708, USA
[b] National Bureau of Economic Research, Cambridge, MA 02138, USA
[c] Purdue University, 610 Purdue Hall, West Lafayette, IN 47906, USA

ARTICLE INFO

Article history:
Received 17 January 2018
Revised 12 August 2020
Accepted 22 August 2020
Available online 21 April 2021

JEL classification:
G11
G14
G23
J24

Keywords:
Factors
Factor selection
Variable selection
Bootstrap
Data mining
Orthogonalization
Multiple testing
Predictive regressions
Fama–MacBeth
GRS

ABSTRACT

Identifying the factors that drive the cross-section of expected returns is challenging for at least three reasons. First, the choice of testing approach (time series versus cross-sectional) will deliver different sets of factors. Second, varying test portfolio sorts changes the importance of candidate factors. Finally, given the hundreds of factors that have been proposed, test multiplicity must be dealt with. We propose a new method that makes measured progress in addressing these key challenges. We apply our method in a panel regression setting and shed some light on the puzzling empirical result that the market factor drives the bulk of the variance of stock returns, but is often knocked out in cross-sectional tests. In our setup, the market factor is not eliminated. Further, we bypass arbitrary portfolio sorts and instead execute our tests on individual stocks with no loss in power. Finally, our bootstrap implementation, which allows us to impose the null hypothesis of no cross-sectional explanatory power, naturally controls for the multiple testing problem.

1. Introduction

What factors explain the cross-section and time series of expected returns? Hundreds of papers propose factors. Numerous new papers purport to help researchers choose a set of factors. However, many challenges remain. The selected set of factors will depend on the test portfolios. Identification is also influenced by the approach, time series versus cross-sectional. Finally, given the large number of candidate factors, some could just be "lucky," working only because of test multiplicity.

To motivate our paper, consider a new factor that debuted in 1965 and presents 45 years of out-of-sample evidence. Using a selection technique such as (Fama and French, 2018), this powerful factor is "chosen" given that

[☆] Previously circulated under the titles "How many factors?" and "Incremental factors." We appreciate the comments of Theodosis Athanasiadis, Charles Clarke, Tony Cookson, Gene Fama, Thomas Flury, Ken French, Bryan Kelly, Hagen Kim, Stefan Nagel, Marco Rossi, and Michael Stutzer, the Editor, Bill Schwert, and three anonymous referees as well as seminar participants at Wharton Jacobs Levy Conference, New York, Society for Quantitative Analysis, New York, Man Quant Conference, New York, the 2016 AFA meetings, and seminars at Texas A&M University, Baylor University, and UC Boulder. We thank Gene Fama, Ken French, Bryan Kelly, and Lu Zhang for sharing their factor returns data. We thank Kay Jaitly for editorial assistance.

[*] Corresponding author.
E-mail address: liu2746@purdue.edu (Y. Liu).

https://doi.org/10.1016/j.jfineco.2021.04.014
0304-405X/© 2021 Elsevier B.V. All rights reserved.

it has an average excess return from 1965 to 2019 of 20% per year. The factor is the excess return on Berkshire Hathaway (BRK) stock.[1] Obviously, this is not an asset pricing factor as it has no ability to explain the cross-section of expected returns. We hope our point is clear: certain methods give substantial weights to high Sharpe ratio factors, such as BRK, that have little to do with the cross-section of expected returns.

Our challenge is threefold. First, we want to gain additional insight into the time series versus cross-sectional aspects of asset pricing tests. For example, why is it that the market factor is by far the most important time series driver of returns (and the first principal component for almost any portfolio formation), yet the market factor is knocked out, or worse, changes sign when combined with other factors? The market factor has obvious theoretical appeal, yet its empirical success is limited.

Second, it is well known that asset pricing results are highly dependent on the portfolio formation technique. For example, the popular HML and SMB factors could work well for 25 portfolios sorted by size and value, yet not so well in portfolios sorted by industry. Ideally, portfolios are not used in the testing, but we all know that portfolios have the advantage of reducing noise.

Third, with hundreds of factors, we need to directly confront the multiple testing problem. In a typical setting for factor tests, we have a set of pre-determined factors and another set of candidate factors. We want to know how well a candidate factor has to perform, taking luck into account, to be declared useful and grouped with pre-determined factors.

Our paper addresses all of these challenges and makes some measured progress on each. We present a step-wise model selection method that is specifically designed for panel regression factor tests and can be easily adapted to alternative testing frameworks. We present a bootstrap method to select additional factors where, under the null, the candidate factor has zero cross-sectional explanatory power. While the bootstrap method is not new, our particular implementation is new and useful for identifying incremental factors.[2]

The main idea behind our bootstrapping approach is the enforcement of the null hypothesis to the data. We show that this enforcement depends on the particular regression model. In panel regressions, we show that enforcing the null hypothesis amounts to adjusting candidate factor returns such that the risk premium of the adjusted factor is completely explained by its exposure to pre-determined factors. As a result, adding the adjusted factor to a regression with pre-determined factors does not change the regression intercept, hence, there is no reduction in alpha, a requirement of the null hypothesis. Our bootstrapping approach then follows White (2000) to control for multiple testing.

We are also able to apply our method to individual stocks, thus avoiding arbitrary formation of portfolios. Our simulation evidence suggests that we do not lose power compared to the portfolio approach. Importantly, our use of individual stocks makes panel regressions the appropriate regression model because, compared to cross-sectional regressions, the impact of extreme observations and potential model mis-specifications are likely mitigated for panel regressions.

Finally, we attempt to gain some insight regarding the market factor, which has a strong theoretical appeal beginning with Sharpe (1964). Even Fama and French (1993) retain the market factor in their analysis, although it fails to explain the cross-section of expected returns in their tests. Our analysis suggests that the market factor is the dominant factor for individual stocks because it substantially reduces pricing errors. We find that the next generation of factors (SMB, HML) is also useful, but of second-order importance. We show that the inference is sensitive to how we weight the pricing errors. For example, if pricing errors are equally weighted, the Fama and French (1993) three-factor model does well. Finally, we explain the strong performance of the market factor in our framework.

Our paper adds to the recent literature on the multidimensionality of the cross-section of expected returns. Harvey et al. (2016) present 316 factors discovered by academia and provide a multiple-testing framework to adjust for data mining. Green et al. (2017) study more than 330 return-predictive signals, which are mainly accounting based, and show large diversification benefits by suitably combining these signals. Hou et al. (2020) study a large collection of anomalies and illustrate the p-hacking (i.e., manipulation of data or estimation methods in a way that generates desired *p*-values) concern through anomaly replication. They argue, following Harvey (2017), that many of the anomalies have been p-hacked. McLean and Pontiff (2016) use an out-of-sample approach to study the post publication bias of discovered anomalies. The overall finding of this literature is that many discovered factors are likely false. But how many factors are useful factors?

Our approach follows an idea similar to one proposed by several predecessor papers in the asset pricing literature that rely on nonparametric bootstrapping, including Ferson and Foerster (1994), Kosowski et al. (2006), Ludvigson and Ng (2009), Fama and French (2010), and Chen et al. (2010). More specifically, finance applications that are related to the White (2000) bootstrap approach include Sullivan et al. (1999) on technical analysis and Fama and French (2010) on mutual fund performance. Our contribution is to show how the bootstrap idea can be embedded in factor tests. In particular, we show how the null hypotheses corresponding to different forms of factor tests can be imposed for a given data set, allowing the use of bootstrapping to nonparametrically control for multiple testing in the selection of useful factors.

[1] In particular, Fama and French's (2018) use of the spanning tests in Barillas and Shanken (2017) will find a large alpha for BRK's return as the left-side portfolio, suggesting the inclusion of BRK into the benchmark factor model. Momentum is another example. Even Fama and French (2015a) augment their five-factor model with momentum because of its high Sharpe ratio and high alpha in spanning regressions.

[2] Recent methods propose a Lasso-type penalty in selecting factors. However, the tuning parameter in these methods serves the role of a multiple-testing adjustment and is non-trivially determined. In contrast, our approach aims at controlling the family-wise error rate, which can be viewed as the multiple-testing adjusted *p*-value. It thus leads to cleaner statistical decision making.

Our empirical analysis focuses on panel regressions that test risk factors with a large cross-section of test assets and is hence related to the well-known Gibbons, Ross, and Shanken (GRS, Gibbons et al., 1989) test. Given that our application features a large cross-section relative to the time series dimension, several issues with the GRS approach are likely insurmountable. First, it is difficult to reliably estimate the residual covariance matrix when the cross-sectional dimension is high.[3] A related issue is that the GRS statistic becomes unusually large with a large cross-section of assets (i.e., the Sharpe ratio for the ex post efficient frontier is unusually large), leading us to reject most factor models. Finally, the GRS approach implies unrealistic (short) positions for certain assets in the ex post mean-variance efficient portfolio (Fama and French, 2015b), casting doubt on its economic interpretation. We propose some alternative test statistics that are motivated by the GRS statistic, but are less reliant on the estimation of the full residual covariance matrix. They are also more interpretable from an economic standpoint in that they try to capture investment opportunities that are financially relevant to investors. We show how to make inference based on these test statistics through a bootstrapping procedure.

Our framework also provides a more intuitive approach than the GRS framework to sort through the myriad of factors that have been proposed over the past 50 years. Suppose we are comparing two candidate factor models: a baseline model M and augmented model M_+. Although researchers frequently use the difference between the GRS statistics generated by M and M_+ to gauge the incremental contribution of M_+,[4] it is not clear how the inference would work. The core of the issue is that M and M_+ are constructed under two different null hypotheses, which are not compatible with each other under the GRS framework. In contrast, the test statistics proposed by our framework allow us to set M as the null model and then evaluate the incremental contribution of M_+. Another benefit of this formulation, given the importance of controlling for multiple tests in the literature on the cross-section of expected stock returns, is that we are able to reevaluate the significance of each M_+ if multiple M_+s have been tried. We achieve this by testing the significance of the max statistic.[5]

We do not solve all the problems, nor do we provide a panacea for asset pricing tests. We believe our technique is a modest innovation that could be particularly useful for factor selection and, more generally, in regressor selection. We believe that using individual stocks provides an opportunity to make apples-to-apples comparisons of models. Currently, it is hard to interpret the literature when one set of authors presents evidence in favor of their five-factor model with one set of portfolios and another set of authors presents evidence in favor of their five-factor model based on a different set of portfolios.

Our paper is organized as follows. In the second section, we present our testing framework. In the third section, we apply our method to the selection of risk factors. We offer insights on both tests based on traditional portfolio sorts as well as individual assets. Some concluding remarks are offered in the final section.

2. Method

We present our model within the context of panel regressions, which corresponds to our main application. Adaptations of our model to predictive regressions and Fama–MacBeth cross-sectional regressions are presented in the online appendix, Section ID, which also contains details of the implementation of our approach. The impact of extreme observations and potential model mis-specifications of individual stock returns is likely mitigated in panel regressions as compared to cross-sectional regressions (we discuss this issue in greater detail later). Thus, our preferred model with individual stocks as test assets is a panel regression framework.

2.1. Panel regression models

Our method can be applied to panel regression models commonly used in asset pricing tests, where asset returns are regressed on a set of common factors. Our goal is to both select candidate factors and to disentangle the time series and cross-sectional explanatory power of the candidates.

We start by writing down a time series regression model,

$$R_{it} - R_{ft} = a_i + \sum_{j=1}^{K} b_{ij} f_{jt} + \epsilon_{it}, i = 1, \dots, N, \tag{1}$$

in which the time series of excess returns $R_{it} - R_{ft}$ is projected onto K contemporaneous factor returns, f_{it}. Factor returns are long-short strategy returns corresponding to zero-cost investments. If the set of factors is mean-variance efficient (or, equivalently, if the corresponding beta-pricing model is true), the cross-section of regression intercepts should be indistinguishable from zero. This constitutes the testable hypothesis for the classic GRS test.

The GRS test is widely applied in empirical asset pricing, but several issues hinder further applications of this test, or time series tests in general. First, the GRS test almost always rejects. This means that almost no model can successfully explain the cross-section of expected returns. As a result, most researchers use the GRS test statistic as a heuristic measure for model performance (see, e.g., Fama and French, 2015a). For instance, if model A generates a smaller GRS statistic than model B, we would take model A as the "better" model, although neither model survives the GRS test. But does model A "significantly" outperform B? The original GRS test cannot answer this question because the overall null of the test is that all intercepts are strictly equal to zero. When two competing models both generate intercepts that are not at zero, the GRS

[3] For example, it is impossible to estimate the covariance matrix if the cross-sectional dimension N is larger than the time series dimension T and we do not assume a special form of the covariance matrix. In our framework, $N < T$ is not required.

[4] See, e.g., Fama and French (2015a,b).

[5] See Politis and Romano (1994), Sullivan et al. (1999), White (2000), and Romano and Wolf (2005).

test is not designed to measure the relative performance of the two models.

Our method provides a potential solution to this model comparison problem. In particular, for two models that are nested, it allows us to measure the incremental contribution of the bigger model relative to the smaller one, even if both models fail to meet the GRS null hypothesis. Second, the inference of the GRS test, which is based on asymptotic approximations for a small N (cross-sectional dimension) and a large T (time series dimension), can be problematic. For instance, with a small N, MacKinlay (1987) shows that the test tends to have low power when the time series sample size is small. Affleck-Graves and McDonald (1990) show that nonnormalities in asset returns can severely distort the GRS test's size and/or power. Our method relies on bootstrapped simulations and should be robust to small-sample or nonnormality distortions. In fact, bootstrap-based resampling techniques are often recommended to mitigate these sources of bias. Finally, for our application with a large N and a relatively small T, the inference for the GRS test does not work.

We try to overcome the aforementioned shortcomings in the GRS test by deploying our bootstrap framework. In particular, we orthogonalize factor returns such that the orthogonalized factors do not impact the cross-section of expected returns.[6] This absence of impact on the cross-section constitutes our null hypothesis. Under this null, we bootstrap to obtain the empirical distribution of the cross-section of pricing errors. We then compare the realized (i.e., based on the real data) cross-section of pricing errors generated under the original factor to this empirical distribution to provide inference on the factor's significance. We describe our panel regression method as follows.

Without loss of generality, suppose we only have one factor (e.g., the excess return on the market $f_{1t} = R_{mt} - R_{ft}$) on the right-hand side of (1). Taking unconditional expectations on both sides of (1), we have

$$E(R_{it} - R_{ft}) = a_i + b_{i1}E(f_{1t}). \tag{2}$$

The mean excess return of the asset can be decomposed into two parts. The first part is the time series regression intercept (i.e., a_i), and the second part is the product of the time series regression slope and the average factor return (i.e., $b_{i1}E(f_{1t})$).

For the one-factor model to work, we need $a_i = 0$ across all assets.[7] Imposing this condition in (2), we have $b_{i1}E(f_{1t}) = E(R_{it} - R_{ft})$. Intuitively, the cross-section of $b_{i1}E(f_{1t})$ needs to line up with the cross-section of expected asset returns (i.e., $E(R_{it} - R_{ft})$) to fully absorb the intercepts in time series regressions. This condition is not easy to satisfy in time series regressions because the cross-section of risk loadings (i.e., the b_i) is determined by individual time series regressions. The risk loadings might happen to line up with the cross-section of expected returns, thereby making the one-factor model work, or they might not. This suggests the possibility that some factors could generate a large time series regression R^2, but contribute little to explaining the cross-section of expected returns.

Another important observation from (2) is that, on the one hand, setting $E(f_{1t}) = 0$, factor f_{1t} has exactly zero impact on the cross-section of expected asset returns. Indeed, if $E(f_{1t}) = 0$, the cross-section of intercepts from time series regressions (i.e., the a_i) exactly equals the cross-section of average excess returns (i.e., $E(R_{it} - R_{ft})$) that the factor model is supposed to help explain in the first place. On the other hand, the factor mean of zero does not matter for time series regressions. Both the regression R^2 and the slope coefficient (i.e., b_{i1}) are preserved.

The previous discussion motivates our test design. For the one-factor model, we define a "pseudo" factor, $\tilde{f}_{1t}$, by subtracting the in-sample mean of f_{1t} from its time series. This demeaned factor maintains all the time series explanatory power of f_{1t}, but has no role in explaining the cross-section of expected returns. With the pseudo factor, we bootstrap (i.e., resample the time periods) to obtain the distribution of a statistic that summarizes the cross-section of pricing errors (i.e., regression intercepts). Candidate statistics include mean/median absolute intercepts, mean-squared intercepts, and absolute t-statistics. We then compare the realized statistic for the original factor (i.e., f_{1t}) to this bootstrapped distribution.

Our method generalizes straightforwardly to the situation when we have multiple factors. Suppose we have K pre selected factors and we want to test the $(K+1)$th factor. We first project the $(K+1)$th factor onto the preselected factors through a time series regression. We then define the new pseudo factor by subtracting the regression intercept from the $(K+1)$th factor. This is analogous to the previous one-factor model example. In the one-factor model, demeaning is equivalent to projecting the factor onto a constant.

We now explore in more detail how our method works when there are multiple factors. With one pre-selected factor, f_{1t}, in the baseline model, the regression equation for asset i is

$$R_{it} - R_{ft} = a_i + b_{i1}f_{1t} + e_{it}. \tag{3}$$

We can add another factor, f_{2t}, to the baseline model and denote the augmented model as

$$R_{it} - R_{ft} = a_i^* + b_{i1}^* f_{1t} + b_{i2}^* f_{2t} + e_{it}^*. \tag{4}$$

If f_{2t} was helpful in explaining the cross-section of expected returns compared to a_i, then a_i^* should be closer to zero. We therefore want to compare a_i^* with a_i. In general, $a_i \neq a_i^*$. Our goal is to adjust f_{2t} such that the adjusted f_{2t} (denoted as f_{2t}^*) guarantees $a_i = a_i^*$, that is, the regression intercept under the augmented model is the same as the intercept under the baseline model. In particular, let the regression equation that projects f_{2t} onto f_{1t} be

$$f_{2t} = \delta_0 + \delta_1 f_{1t} + \varepsilon_t, \tag{5}$$

and define f_{2t}^* as

$$f_{2t}^* \equiv f_{2t} - \delta_0 = \delta_1 f_{1t} + \varepsilon_t. \tag{6}$$

[6] More precisely, our method ensures that the orthogonalized factors have zero impact on the cross-section of expected returns *unconditionally* because panel regression models with constant risk loadings focus on unconditional asset returns. In later sections, we discuss extensions of our framework to cope with conditional expected returns.

[7] We explain in greater detail what we mean by for a model to "work" in the next section.

Thus defined, f^*_{2t} when substituting f_{2t} in (4) ensures that $a^*_i = a_i$.[8] To see this, we replace f^*_{2t} with f_{2t} in (4) and rewrite the regression equation as

$$R_{it} - R_{ft} = a^*_i + b^*_{i1} f_{1t} + b^*_{i2}(\delta_1 f_{1t} + \varepsilon_t) + e^*_{it}, \tag{7}$$

$$= a^*_i + (b^*_{i1} + b^*_{i2}\delta_1) f_{1t} + \underbrace{(b^*_{i2}\varepsilon_t + e^*_{it})}_{u^*_{it}}. \tag{8}$$

By construction, both ε_t and e^*_{it} are orthogonal to f_{1t} and a vector of ones. Hence, by treating $u^*_{it} = b^*_{i2}\varepsilon_t + e^*_{it}$ as the new regression residual and by comparing (8) and (3), we have

$$a^*_i = a_i, \quad b^*_{i1} + b^*_{i2}\delta_1 = b_{i1}. \tag{9}$$

Our adjustment makes economic sense. Taking unconditional expectations on both sides of (6), we have

$$E(f^*_{2t}) = \delta_1 E(f_{1t}). \tag{10}$$

Therefore, the adjusted factor, f^*_{2t}, is absorbed by the preselected factor f_{1t} in the sense that its premium is completely explained by its exposure to the pre-selected factor. When this happens, the adjusted factor has a zero incremental impact on the cross-section of expected returns. In the meantime, it has perfect time series correlation with the original factor in sample and has the same time series correlation with the pre selected variable as the original factor. Hence, the adjusted factor preserves the time series properties of the original factor aside from the mean.

With this pseudo factor, we bootstrap (i.e., resample the time periods) to generate the distribution of pricing errors. In this step, unlike the one-factor case, for both the original regression and the bootstrapped regressions based on the pseudo factor, we always keep the original K factors in the model. This way, our test captures the incremental contribution of the candidate factor and can be characterized as a *stepwise model selection method*.

2.2. Alternative regression frameworks

We can straightforwardly adapt our technique to different settings such as predictive regressions and Fama and MacBeth (1973) cross-sectional regressions. In the online appendix in Section ID.1, we use the predictive regression setting to discuss other details of our bootstrap implementation. We then recast our model within the Fama–MacBeth regression framework in Section ID.2 of the online appendix.

Our main application of this method is to individual stocks. We believe that panel regressions should be preferred to cross-sectional regressions (i.e., Fama–MacBeth regressions) when using individual stocks. The reason is that panel regressions allow asset-specific intercepts to absorb idiosyncrasies that are unrelated to the regressors of interest, thereby mitigating the impact of extreme observations and model misspecification. Indeed, our simulation study shows that panel regressions are more powerful than cross-sectional regressions.[9]

2.3. Discussion

2.3.1. Bootstrapped hypothesis testing

Across the three different scenarios (i.e., predictive regressions, panel regressions, and Fama–MacBeth cross-sectional regressions), our orthogonalization works by adjusting the right-hand side so they appear irrelevant in sample; that is, they achieve the null hypothesis in sample. Under the null hypothesis, one can resample to obtain the bootstrapped distribution of a test statistic to perform hypothesis testing.

Our approach is closely related to existing papers in asset pricing that use a nonparametric bootstrap procedure to make inference. For example, Kosowski et al. (2006) and Fama and French (2010) impose the null of zero alphas and use bootstrapping to make inferences about fund performance. Ferson and Foerster (1994) use bootstrapping in a simulation environment to generate alternative draws from the data-generating process. Ludvigson and Ng (2009) apply bootstrapping to obtain generalized standard errors that account for the uncertainty in their first-stage factor model estimation. We use bootstrapping in a way that is similar to Kosowski et al. (2006) and Fama and French (2010) in that we also attempt to force the null hypothesis to hold in sample and bootstrap to perform hypothesis testing. The novel element of our use of bootstrapping is the construction of the null hypothesis, which is specific to our tests of factors and varies across different regression frameworks. We discuss bootstrapped hypothesis testing and the related literature in the online appendix, Section IE.1.

2.3.2. “Useful” factors

We consider a risk factor “useful” if it helps to explain the cross-section of expected returns, that is, it helps reduce panel regression alphas. We provide the details in online appendix, Section IE.2.

2.3.3. Multiple hypothesis testing

Our framework is designed to curb the proliferation of factors analyzed in recent papers by Harvey et al. (2016), Green et al. (2017), and Hou et al. (2020). In particular, we set up factor tests in a way that allows us to apply the reality bootstrap of White (2000) and Romano and Wolf (2005), which aims to control the family-wise error rate (FWER), that is, the probability of making at least one false discovery among a number of tests. In the online appendix, Section IE.3, we discuss related papers on multiple hypothesis testing.

[8] Our method is context specific. In predictive regressions (online appendix, Section ID.1), x_{2t} is orthogonalized with respect to x_{1t}. In contrast, we simply adjust the mean of x_{2t} in panel regressions.

[9] Extreme observations can also impact the two methods differentially. For the individual stock return data, extreme observations exist and tend to have a large influence on cross-sectional regressions. In contrast, each individual stock is weighted equally (or value weighted, implying even smaller weights on small stocks that tend to generate extreme observations) in panel regressions, leading to a much smaller impact by extreme observations.

2.3.4. Sequential tests

We discuss the pros and cons of our step-wise model selection technique compared to alternative approaches in online appendix, Section IE.4. Note that whether our approach is sequential or not does not affect its validity at each stage of our test (i.e., does the best factor among a group of factors survive the multiple-testing adjusted statistical threshold, possibly with the presence of preexisting factors)? Exemplar papers in the subsequent literature that adopt our testing framework include those by Lin, Palazzo, and Yang (2020), Chen et al. (2019), and Croce et al. (2019). These papers take as given a preselected set of factors and use our approach to evaluate the incremental contributions of their proposed factors.

2.3.5. Related literature

We provide an extensive discussion of the related literature. In particular, we discuss the connection between our approach and the GRS test, reconcile the difference between our tests and the popular (Barillas and Shanken, 2017) approach, and relate our framework to the stochastic discount factor (SDF) literature. We provide details of our discussion in the online appendix, Section IE.5.

3. Application: identifying useful factors

We now apply our panel regression model to the selection of risk factors. We incrementally select risk factors that are helpful in explaining the cross-section of expected returns, while controlling for test multiplicity.

3.1. Candidate risk factors

In principle, we can apply our method to the grand task of sorting all the risk factors that have been proposed. One attractive feature of our method is that it allows the number of risk factors to be larger than the number of test portfolios, which is infeasible in conventional multiple-regression models. We do not pursue this in the current paper, however, but instead focus on a select group of prominent risk factors. The choice of the test portfolios is a major confounding issue. Different test portfolios lead to different results. In contrast, individual stocks avoid arbitrary portfolio construction. We apply our method to both popular portfolio sorts as well as individual stocks.

In particular, we apply our panel regression method to 14 risk factors proposed by Fama and French (2015a), Frazzini and Pedersen (2014), Novy-Marx (2013), Pástor and Stambaugh (2003), Carhart (1997), Asness et al. (2019), Hou et al. (2015), Harvey and Siddique (2000), and Herskovic et al. (2016).[10]

We first provide acronyms for factors. Fama and French (2015a) add profitability (*rmw*) and investment (*cma*) to the three-factor model of Fama and French (1993), which starts with the Sharpe (1964) market (*mkt*) factor and is augmented with size (*smb*) and book-to-market (*hml*) as pricing factors. Hou et al. (2015) propose similar profitability (*roe*) and investment (*ia*) factors. Other factors include betting againist beta (*bab*) from Frazzini and Pedersen (2014), gross profitability (*gp*) from Novy-Marx (2013), Pástor and Stambaugh liquidity (*psl*) from Pástor and Stambaugh (2003), momentum (*mom*) from Carhart (1997), quality minus junk (*qmj*) from Asness et al. (2019), coskewness (*skew*) from Harvey and Siddique (2000), and common idiosyncratic volatility (*civ*) from Herskovic et al. (2016). We treat these 14 factors as candidate risk factors and incrementally select the group of factors.

Table 1 presents the summary statistics for the 14 main factors on which we focus. The 14 risk factors generate sizable long-short strategy returns. Nine of the strategy returns generate t-ratios above 3.0, which is a potential new cutoff proposed in Harvey et al. (2016). The correlation matrix suggests a clustering of some of the factors. First is the "value" group, consisting of the book-to-market ratio (*hml*), the Fama and French (2015a) investment factor (*cma*), and the Hou et al. (2015) investment factor (*ia*). Second is a "profitability" group, consisting of the Fama and French (2015a) profitability factor (*rmw*), the Hou et al. (2015) profitability factor (*roe*), and the Asness et al. (2019) quality-minus-junk factor (*qmj*). For example, *cma* and *ia* have a correlation of 0.90, and *rmw* and *qmj* have a correlation of 0.76.

To examine the sensitivity of our results to the set of factors we study, we also consider an extended list of 56 factors (including the 14 main factors we have just described) that are used by Ehsani and Linnainmaa (2021).[11] We refer to Ehsani and Linnainmaa (2021) for details of their factors.

3.2. Test statistics

We focus on test statistics that are both economically and statistically sound. Intuitively, a good test statistic in our context should be able to reflect the difference, in explaining the cross-section of expected returns, between a baseline model and an augmented model that adds one additional variable to the baseline model. For the panel regression model, let $\{a_i\}_{i=1}^N$ and $\{a_i^+\}_{i=1}^N$ be the cross-section of regression intercepts for the baseline model and the augmented model, respectively. Let $\{s_i\}_{i=1}^N$ be the cross-section of standard errors for regression intercepts under the baseline model. Our first test statistic is given by

$$SI_{ew}^m \equiv \frac{\frac{1}{N}\sum_{i=1}^N (|a_i^+| - |a_i|)/s_i}{\frac{1}{N}\sum_{i=1}^N |a_i|/s_i},$$

where SI denotes scaled intercept, ew denotes equal weighting, and m denotes mean. Intuitively, SI_{ew}^m measures the percentage difference in the absolute regression intercepts scaled by the standard error for the regression inter-

[10] The factors of Fama and French (2015a), Hou et al. (2015), Harvey and Siddique (2000), and Herskovic et al. (2016) are provided by the authors. The factors from the rest of the papers are obtained from the authors' web pages. Across the 14 factors, the Pástor and Stambaugh (2003) liquidity factor has the shortest length (i.e., January 1968–December 2012). We therefore focus on the January 1968 to December 2012 period so that all factors have the same sampling period.

[11] Ehsani and Linnainmaa (2021) consider 52 factors, among which 10 overlap with the 14 main factors we consider. Merging the two sets of factors, we have 56 distinct factors.

Table 1
Summary statistics for monthly factor returns, January 1968–December 2012.

This table shows summary statistics on factors. We report the mean annual returns for the five risk factors in Fama and French (2015a) (i.e., excess market return (*mkt*), size (*smb*), book-to-market (*hml*), profitability (*rmw*), and investment (*cma*)), betting against beta (*bab*) in Frazzini and Pedersen (2014), gross profitability (*gp*) in Novy-Marx (2013), Pástor and Stambaugh liquidity (*psl*) in Pástor and Stambaugh (2003), momentum (*mom*) in Carhart (1997), quality minus junk (*qmj*) in Asness et al. (2019), investment (*ia*) and profitability (*roe*) in Hou et al. (2015), coskewness (*skew*) in Harvey and Siddique (2000), and common idiosyncratic volatility (*civ*) in Herskovic et al. (2016). We also report the correlation matrix for factor returns. The sample period is from January 1968 to December 2012. High correlations are emphasized in bold.

Panel A: Factor returns

	mkt	*smb*	*hml*	*mom*	*skew*	*psl*	*roe*	*ia*	*qmj*	*bab*	*gp*	*cma*	*rmw*	*civ*
Mean	0.052	0.022	0.048	0.081	0.024	0.055	0.068	0.057	0.048	0.105	0.039	0.047	0.033	0.060
t-stat	[2.17]	[1.32]	[3.08]	[3.54]	[1.84]	[2.99]	[5.09]	[5.76]	[3.74]	[5.98]	[3.24]	[4.44]	[2.92]	[3.48]

Panel B: Factor correlation matrix

	mkt	*smb*	*hml*	*mom*	*skew*	*psl*	*roe*	*ia*	*qmj*	*bab*	*gp*	*cma*	*rmw*	*civ*
mkt	1.00													
smb	0.30	1.00												
hml	−0.32	−0.24	1.00											
mom	−0.14	−0.03	−0.15	1.00										
skew	−0.02	−0.05	0.23	0.03	1.00									
psl	−0.05	−0.04	0.03	−0.03	0.10	1.00								
roe	−0.19	−0.39	−0.11	0.51	0.19	−0.06	1.00							
ia	−0.39	−0.26	**0.69**	0.04	0.15	0.02	0.04	1.00						
qmj	−0.54	−0.54	0.02	0.26	0.13	0.03	**0.68**	0.15	1.00					
bab	−0.09	−0.07	0.40	0.18	0.24	0.06	0.25	0.35	0.19	1.00				
gp	0.08	0.06	−0.34	0.01	−0.01	−0.03	0.34	−0.26	0.45	−0.11	1.00			
cma	−0.41	−0.16	**0.71**	0.01	0.05	0.03	−0.10	**0.90**	0.07	0.32	−0.34	1.00		
rmw	−0.21	−0.42	0.11	0.10	0.27	0.03	**0.68**	0.05	**0.76**	0.26	0.49	−0.08	1.00	
civ	0.17	0.27	0.13	−0.18	0.04	0.05	−0.26	−0.00	−0.28	0.11	−0.00	0.04	−0.10	1.00

cept under the baseline model.[12] We would expect SI^{m}_{ew} to be negative (i.e., pricing errors are lower) if the augmented model is better than the baseline model. The significance of the improvement is evaluated against the bootstrapped empirical distribution generated under the null hypothesis that the additional variable in the augmented model has zero incremental contribution in explaining the cross-section of expected returns.

Whereas SI^{m}_{ew} calculates the percentage difference in the scaled mean absolute intercept, it may not be robust to extreme observations in the cross-section, especially when we use individual stocks as test assets. We therefore also consider a robust version that calculates the percentage difference in the scaled median absolute intercept, that is,

$$SI^{med}_{ew} \equiv \frac{med(\{|a_i^+|/s_i\}_{i=1}^N) - med(\{|a_i|/s_i\}_{i=1}^N)}{med(\{|a_i|/s_i\}_{i=1}^N)},$$

where *med* denotes the median of a group of variables and is denoted by the superscript *med*.

One key assumption for the validity of our test statistic is that the cross-sectionally averaged $|a_i^+|$ should be smaller than the cross-sectionally averaged $|a_i|$ if the additional factor in the augmented model is useful, that is, if the augmented model does a better job of explaining the cross-section of expected returns. At the individual asset level, $|a_i^+|$ will be smaller than $|a_i|$ in population (i.e., we have sufficiently long factor and return time series) if the augmented model is the true underlying factor model. Indeed, as we discussed previously, if the augmented model is correctly specified, then a_i^+ will be zero in population, which is no greater than $|a_i|$ under the (incorrectly specified) baseline model.

Several reasons support our consideration of the scaled intercept instead of the original intercept. First, in a time series regression model, by thinking of the fitted combination of zero-cost portfolios (that is, factor proxies) as a benchmark index, the scaled intercept is closely related to the *information ratio* of the strategy that takes a long position in the test asset and a short position in the benchmark index.[13] When test assets are not diversified portfolios, the information ratio is a better scaled metric to gauge the economic significance of the investment strategy. This is similar to the use of the *t*-statistic instead of Jensen's alpha in performance evaluation. The *t*-statistic of alpha, not alpha itself, tells us how "abnormal" a fund manager's returns are.

Second, the use of the scaled intercept takes the heterogeneity in return volatilities into account. Suppose two stocks generate the same regression intercept when fitting a factor model. The degree of mispricing by the factor model, as measured by the absolute value of the regression intercept, should be higher for the less noisy stock. In other words, we should assign less weight to the noisier stocks in our panel regression model. This is particularly important when we consider individual stocks as test assets because of the large amount of heterogeneity in return volatility.

Finally, as discussed in online appendix, Section IE.1, our use of the scaled intercept is consistent with the second principle for the bootstrap hypothesis testing of Hall and Wilson (1991). In fact, scaling the intercept by the stan-

[12] Alternatively, we could use the appraisal ratio or information ratio (i.e., alpha divided by the residual standard deviation) to construct our test statistic.

[13] See Treynor and Black (1973).

dard error is exactly what they recommend to obtain pivotal statistics.

Another important feature of our test statistics is that we scale the intercepts of the baseline model and the augmented model by the same standard error, that is, the standard error of the estimate of the intercept under the baseline model. This ensures that our test statistics are exactly zero when the null hypothesis, i.e., the candidate factor has zero incremental contribution to explain the cross-section of expected returns, is forced to exactly hold in sample for our procedure. This might not hold under alternative scaling schemes. For example, we could propose the use of the standard errors corresponding to the baseline model and the augmented model to separately scale the regression intercepts under the two models. This does not work in our setup, however, as the orthogonalized candidate factor (e.g., the demeaned market factor), which is constructed to have a zero impact on the regression intercepts, could have non-negligible impacts on the standard errors. As a result, the test statistic will not equal zero at the null hypothesis because the same intercept is scaled by two different standard errors, making it difficult to disentangle the cross-sectional impact of a candidate factor from its time series impact. Our test statistics allow us to single out the cross-sectional contribution of the candidate factor.

What is the difference between our test and the GRS test? The GRS test hypothesizes that the augmented model is true and evaluates the cross-section of $|a_i^+|$ to test this hypothesis. A failure of the test indicates the rejection of the augmented model as a whole, but tells us little about the significance of individual factors. In contrast, our test hypothesizes that the baseline model is true and uses the reduction in scaled absolute intercepts to evaluate the incremental contribution of the additional factor in the augmented model. As a result, our test is able to tell whether the additional factor is individually significant as a risk factor without having to make a statement about overall model performance. But, given the uncertainty about the underlying true factor model, any given factor model is likely to be misspecified. This is important given the difficulty in having a correctly specified model for a large cross-section of assets.[14] We delve further into this point in the next section.

Another difference is that, instead of using the entire residual covariance matrix to weight the cross-section of regression intercepts as the GRS test does, we use the individual standard errors. This difference might be considered a drawback of our test because GRS would seem more powerful because it uses information in the correlation structure of returns. That is, it allows the use of the residual covariance matrix to construct portfolios. These portfolios are more mean-variance efficient than those based on the tested factors alone, thereby improving test power. In reality, however, the instability in the estimation of the covariance matrix could offset the gain in test power. As discussed in online appendix, Section IE.5, the GRS tangency portfolio often implies economically implausible weights on certain assets, which is problematic for the interpretation of the test. Additionally, when we use individual stocks as test assets, the covariance matrix will be poorly measured because the number of assets in the cross-section is larger than the number of periods in the time series, rendering the GRS test inapplicable.[15] Given these concerns, we believe that our test could have some advantages over the GRS test when applied to a large cross-section of assets. Our test takes the residual volatility for each individual asset into account, while at the same time avoiding the estimation of the large dimensional residual covariance matrix.[16] We provide a detailed comparison of our method and the GRS test toward the end of this section.

Although we focus on the test statistics SI_{ew}^{m} and SI_{ew}^{med}, many other test statistics are feasible. For example, instead of using the scaled intercepts, it is possible to use the original intercepts. Another example is to use value weighting rather than equal weighting of the intercepts. We explore alternative test statistics in later sections.

The fact that our framework allows us to consider a variety of test statistics demonstrates the flexibility of our bootstrap approach. For example, we usually do not have closed-form asymptotic approximations for test statistics that are based on quantiles (e.g., the median). Our bootstrap-based approach allows us to provide inference for test statistics that rely on the median, which is robust to outliers and is especially important for our application to individual stocks.

Instead of using the equally weighted scaled intercepts, Fama and French (2015a) use the equally weighted absolute intercepts (and alternative ways of weighting the intercepts) as the heuristic test statistic to evaluate the performance of their investment and profitability factors. Our framework allows us to make precise statements about the statistical significance of their test statistics. We show in simulation studies that our test statistics are much more powerful than the unscaled intercept-based test statistics.[17] We therefore focus on the two aforementioned test statistics (i.e., SI_{ew}^{m} and SI_{ew}^{med}) and their value-weighted counterpart (i.e., SI_{vw}^{m}), which will be introduced later in our paper.

Barillas and Shanken (2017) propose the use of a quadratic form of the intercepts, which takes into account

[14] See Kan and Robotti (2009) for a similar point.

[15] While dimension reduction techniques for estimating the covariance matrix are available (e.g., Ledoit and Wolf, 2003), their use does not solve the issue of having very large GRS statistics. In addition, we lose the analytical tractability of the GRS test. See online appendix Section IE.5 for detailed discussion on the GRS test.

[16] One reason for the popularity of the GRS test is that its weighting scheme leads to a test statistic whose distribution does not depend on unknown model parameters under the null hypothesis. As a result, researchers can conveniently refer to the standard F distribution to perform hypothesis testing. Under our weighting scheme, the distribution of the test statistic under the null hypothesis will depend on model parameters such as the residual volatilities of individual assets. Fortunately, our bootstrap-based framework provides a convenient way to provide inference.

[17] A simulation study similar to Appendix B shows that our test statistics based on scaled intercepts have higher power than test statistics based on absolute intercepts, with the average improvement in test power around 15% across factors and specifications for factor risk premiums. Our results are available on request.

the correlation of intercepts, to compare candidate factor models. In our application to individual stocks, it becomes infeasible to invert the large dimensional covariance matrix for intercepts. In addition, Affleck-Graves and McDonald (1990) show that the GRS statistic constructed using only the diagonal elements of the error covariance matrix performs better in their simulation studies. We therefore focus on the two test statistics that do not adjust for correlations in intercepts.

We can also interpret our test statistics from an investment perspective. We save such interpretations for later sections, where we contrast our regression results with those under the GRS approach.

3.3. Main results: individual stocks as test assets

Instead of characteristic-sorted portfolios, can we use individual stocks to provide inference? Conventional wisdom says no. Indeed Jensen et al. (1972) and Fama and MacBeth (1973) argue that individual stocks are too noisy to serve as test assets. The GRS test also prohibits the use of individual stocks as the inversion of the large variance-covariance matrix of the return residuals is problematic. Subsequent researchers follow these suggestions and use popular portfolios, in particular the Fama–French 25 portfolios, to test asset pricing factors.

A counter argument is that the use of portfolios could introduce bias and inefficiency into asset pricing tests. Avramov and Chordia (2006) show that the asset pricing implications differ a lot when we use single securities instead of portfolios. Ang et al. (2016) argue that the larger dispersion in beta resulting from the use of individual stocks can potentially enhance the power of the test.[18] Lewellen et al. (2010) suggest the use of a large number of assets instead of a small number of portfolios to judge model performance.

We believe that the single most important reason for considering individual stocks is that they provide a cleaner test of risk factors. Tests based on characteristics-sorted portfolios are likely to be biased towards identifying the risk factors that are constructed using the same set of test portfolios (Ferson and Harvey, 1999; Berk, 2000; Pukthuanthong et al., 2019). The use of individual stocks guards against the data-snooping bias induced by portfolio-based asset pricing tests, as shown in Lo and MacKinlay (1990). Additionally, if we rephrase the argument in Jensen et al. (1972) and Fama and MacBeth (1973) to say that the high level of noise in individual stocks renders most tests inefficient, then the problem is not necessarily about individual stocks themselves, but about the lack of statistical tests that are robust to the noise in stocks.[19] Our paper provides such a test. In particular, our framework allows us to make inferences about test statistics that take stock volatilities and market values into account. It therefore mitigates the noise issue for individual stocks (by downweighting small and noisy stocks).

Besides the bias issue, our framework provides a systematic approach to overcome many of the challenges in the use of individual stocks, including an unbalanced panel, extreme observations in the cross-section, cross-sectional dependence (possibly nonparametric), and a large cross-section relative to time series.

3.3.1. A simulation study

We first perform a simulation study to see whether our tests have the statistical power to identify a “true” risk factor, i.e., a risk factor that belongs to the underlying factor model. This is important given the concern that the high level of noise in individual stock returns might render our test powerless. We do not pursue a full-blown simulation study that investigates every aspect of our procedure, but instead focus on the selection of a candidate risk factor that provides additional information to the market factor in explaining the cross-section of expected returns.[20] This is motivated by the fact that the market factor is the single dominating factor that is always selected first in our empirical analysis. It is more interesting to examine the next factor that enters our factor list.

Two features distinguish our simulation study from standard simulation frameworks. First, we take the cross-sectional distribution of factor loadings in the data as given. Second, we bootstrap the realized return residuals to construct the simulated panel of returns. Compared to standard simulation studies that assume a parametric distribution for the factor loadings and/or return residuals, our method brings the simulated data closer to the actual data, and therefore provides a more realistic assessment of test power. Appendix B describes our simulation study in detail.

To benchmark our results against existing methods, we consider the popular approach, which sorts stocks based on estimated betas. In particular, we estimate factor loadings based on a five-year rolling window and construct a long-short portfolio that we hold out of sample for one year. We use the *t*-statistic of the portfolio returns to test the significance of the candidate factor. For comparison purposes, we also consider a simple unconditional beta sorting scheme that first sorts stocks based on their unconditional univariate factor loadings estimated over the entire sample, and then forms the long-short portfolio by having a long position in stocks that are in the top beta decile and a short position in stocks that are in the bottom beta decile.[21]

There are several takeaways from our simulation study, as summarized in online appendix Tables A.2 and A.3. First,

[18] Estimation uncertainty for the estimated betas makes the gain in power using the method in Ang, Liu, and Schwarz (2016) vanish asymptotically. However, in finite samples, the gain in power from using a larger set of test assets is likely to be substantial.

[19] Because of changing firm characteristics, the use of individual stocks could bias the tests against finding certain characteristics (especially characteristics that are not highly persistent). One can adapt our framework to allow for time-varying stock characteristics or study the subsample performance of our current approach. See online appendix Section IG.1 for more detailed discussion.

[20] For example, we do not evaluate the accuracy of the bootstrapped distribution in approximating the true underlying distribution, as is done in White (2000). As another example, we focus on constant factor loadings and do not model the potential dependency between factor loadings and time-varying firm characteristics.

[21] Sorts based on multivariate factor loadings yield similar results as the candidate factors are not highly correlated with the market factor.

compared to either conditional or unconditional beta sorts, our bootstrap-based tests are more powerful. In particular, when the factor risk premium is similar to what we see in the real data, the power of our tests based on the t-statistics is on average (across different factors) about 10% (i.e., percentage points) higher than that based on conditional beta sorts. Additionally, when the number of time periods is about the same as in the real data, the power of our tests is well above 70% across different risk factors. Hence, our tests have power in an absolute sense as well.

There is also a notable difference in the test size across different testing methods. In particular, the two beta sorts we benchmark against have actual sizes that are somewhat below the nominal size. This is consistent with their lower power when we deviate from the null hypothesis. The fact that the two beta sorts are undersized can be explained by the errors-in-variable (EIV) bias (i.e., Shanken, 1992) for the cross-sectional approach over a relatively short time series sample. Our panel regression approach does not suffer from the EIV bias found in cross-sectional regressions. Jegadeesh et al. (2019) propose an EIV correction for the cross-sectional approach with individual stocks. While we do not replicate their simulation results, their Table 3 reports an average test power of around 76% for the Fama–French three-factor model with 684 monthly observations. In comparison, the average test power for our approach in online appendix Table A.3 (with $A = 1.0$, i.e., risk premium is set at the in-sample estimate) is 83% across the five factors with only 480 monthly observations.[22] Hence, our tests perform well in comparison to their recently proposed cross-sectional approach. Nonetheless, to control for the difference in the actual size across tests, we follow the idea in Harvey and Liu (2020a) and report the size-adjusted test power in appendix Table A.4. Our approach still compares favorably with the two beta sorts in terms of test power.

Second, we are not necessarily losing test power by considering individual stocks. In particular, we redo our exercise using Fama–French 25 portfolios and find that the average performance of our tests based on portfolios is similar to that of our tests based on individual stocks. The key assumption for our simulation study is that a two-factor model is the true underlying factor model. In reality, the Fama–French 25 portfolios follow a tight factor structure and therefore are likely to favor factors that are correlated with the Fama–French factors (see Lewellen et al., 2010). In contrast, individual stocks provide a diverse set of assets.

3.3.2. Value weighted test statistic

In addition to the previously mentioned test statistics that rely on equally weighted scaled regression intercepts, we consider a value-weighted version of them. Value weighting makes economic sense. For two stocks that generate the same regression intercept, the mispricing of the factor model should be more economically important for the stock that has a higher market value. Our value-weighted test statistic therefore uses market values to weight the cross-section of scaled intercepts. In particular, let $\{me_{i,t}\}_{t=1}^{T}$ be the time series of market equity for stock i, and let $ME_t = \sum_{i=1}^{N} me_{i,t}$ be the aggregate market equity at time t. The test statistic is given by

$$SI_{vw}^{m} \equiv \frac{\sum_{t=1}^{T}\sum_{i=1}^{N}\frac{me_{i,t}}{ME_t}\times(|a_i^{+}|-|a_i|)/s_i}{\sum_{t=1}^{T}\sum_{i=1}^{N}\frac{me_{i,t}}{ME_t}\times|a_i|/s_i},$$

where vw denotes value weighting. SI_{vw}^{m} calculates the percentage difference in the time-averaged value-weighted level of mispricing between the augmented model and the baseline model. Our value-weighted test statistic takes the time variation in market value into account. Notice that while value-weighted test statistics more accurately reflect the wealth effect experienced by the average investor (see Fama, 1998), they substantially reduce the impact of smaller sized stocks.

3.3.3. Test results with individual stocks

We apply our method to individual stocks. In later sections, we apply our approach to popular test portfolios and offer a detailed comparison between our results based on individual stocks and portfolios. We highlight the potential bias generated by tests based on sorted portfolios.

For our empirical analysis throughout the paper, at each step of our incremental selection procedure, we report both the single-test p-value for each factor and the multiple-test p-value for the minimum test statistic.[23] To be clear, our decisions are based on multiple-test p-values, however, we also report the single-test p-values to highlight the role of test multiplicity.

Tables 2 (equal weighting) and 3 (value weighting) present the results based on individual stocks. For equal-weighted test statistics, we focus on SI_{ew}^{med} to interpret our results since it is more robust to outliers than SI_{ew}^{m}. That said, our results are consistent (i.e., SI_{ew}^{med} vs. SI_{ew}^{m}) in terms of the factors that are identified.

Note that the market factor is significant from both a single-test and a multiple-test perspective. From a single-test perspective, to evaluate the significance of the market factor, we follow our method and orthogonalize the 14 factors so they have a zero impact on the cross-section of expected returns in-sample. We bootstrap to obtain the empirical distributions of the individual test statistics. We then evaluate the realized test statistics against these empirical distributions to provide p-values. As shown in Panel A of Table 2, the bootstrapped 5th percentile of SI_{ew}^{med} for the market factor is -0.095. The interpretation is that bootstrapping under the null, i.e., the market factor has no ability to explain the cross-section, produces a distribution of increments to the intercept. At the 5th percentile, there is a percentage reduction in the median scaled intercept of -9.5%. The actual factor reduces the mean scaled intercept by more than the 5th percentile, so we declare it significant. More precisely, by evaluating the 20.6% reduction against the empirical distribution of SI_{ew}^{med} for the market

[22] Note that Jegadeesh et al.'s (2019) simulation study also differs from ours in terms of the number of stocks in the cross section and how return residuals are drawn. Therefore, our comparison with Jegadeesh et al. (2019) is not apples-to-apples.

[23] The single-test p-values are easily obtained from our bootstrap framework.

Table 2
Individual stocks as test assets, equally weighted scaled intercepts.1

This table shows test results for 14 risk factors using equally weighted individual stocks. (See Table I for the definitions of risk factors.) A stock needs to have at least 36 monthly observations (either in the original or the bootstrapped sample) to enter our tests. The baseline model refers to the model that includes the pre selected risk factors. We focus on the panel regression model described in Section 2.1. The two metrics (i.e., SI^{m}_{ew} and SI^{med}_{ew}), which measure the difference in the equally weighted scaled mean/median absolute regression intercept, are defined in Section 3.2. Bold numbers are associated with the best incremental factor. The 5th percentile and *p*-value for multiple test correspond to the multiple-testing adjusted 5th percentile and the *p*-value for the best incremental factor, respectively.

	Panel A: Baseline = No factor						Panel B: Baseline = *mkt*					
		single test			single test			single test			single test	
Factor	SI^{m}_{ew}	5th-percentile	*p*-value	SI^{med}_{ew}	5th-percentile	*p*-value	SI^{m}_{ew}	5th-percentile	*p*-value	SI^{med}_{ew}	5th-percentile	*p*-value
mkt	−**0.192**	[−0.093]	(0.003)	−**0.206**	[−0.095]	(0.001)						
smb	−0.081	[−0.081]	(0.056)	−0.109	[−0.117]	(0.061)	−**0.041**	[−0.045]	(0.063)	−**0.062**	[−0.052]	(0.032)
hml	0.088	[−0.022]	(0.983)	0.108	[−0.029]	(1.000)	−0.021	[−0.030]	(0.131)	−0.047	[−0.028]	(0.014)
mom	0.091	[−0.034]	(1.000)	0.110	[−0.044]	(1.000)	0.070	[−0.007]	(1.000)	0.089	[−0.012]	(1.000)
skew	−0.008	[−0.031]	(0.278)	−0.002	[−0.034]	(0.478)	−0.004	[−0.009]	(0.167)	−0.003	[−0.013]	(0.319)
psl	0.011	[−0.019]	(0.920)	0.002	[−0.030]	(0.682)	0.001	[−0.004]	(0.409)	−0.003	[−0.012]	(0.237)
roe	0.163	[−0.042]	(0.951)	0.187	[−0.064]	(1.000)	0.142	[−0.019]	(1.000)	0.180	[−0.029]	(1.000)
ia	0.264	[−0.040]	(1.000)	0.291	[−0.048]	(1.000)	0.027	[−0.009]	(0.968)	0.015	[−0.015]	(0.934)
qmj	0.316	[−0.072]	(0.995)	0.358	[−0.090]	(0.998)	0.149	[−0.024]	(0.972)	0.193	[−0.029]	(0.973)
bab	−0.006	[−0.039]	(0.594)	−0.049	[−0.050]	(0.107)	0.018	[−0.010]	(0.983)	−0.014	[−0.017]	(0.181)
gp	0.017	[−0.008]	(0.529)	0.030	[−0.007]	(0.727)	0.023	[−0.005]	(0.961)	0.017	[−0.007]	(0.790)
cma	0.176	[−0.034]	(1.000)	0.199	[−0.035]	(1.000)	−0.012	[−0.013]	(0.057)	−0.031	[−0.019]	(0.027)
rmw	0.116	[−0.011]	(0.986)	0.137	[−0.017]	(0.994)	0.040	[−0.014]	(1.000)	0.048	[−0.020]	(0.975)
civ	−0.096	[−0.044]	(0.023)	−0.130	[−0.062]	(0.031)	−0.018	[−0.018]	(0.052)	−0.049	[−0.030]	(0.021)
		multiple test			multiple test			multiple test			multiple test	
	min	[−**0.109**]	**(0.004)**		[−**0.147**]	**(0.002)**		[−**0.045**]	**(0.071)**		[−**0.057**]	**(0.039)**
	Panel C: Baseline = *mkt*+*smb*						Panel D: Baseline = *mkt* + *smb*+*hml*					
		single test			single test			single test			single test	
Factor	SI^{m}_{ew}	5th-percentile	*p*-value	SI^{med}_{ew}	5th-percentile	*p*-value	SI^{m}_{ew}	5th-percentile	*p*-value	SI^{med}_{ew}	5th-percentile	*p*-value
mkt												
smb												
hml	−**0.017**	[−0.020]	(0.061)	−**0.040**	[−0.025]	(0.011)						
mom	0.055	[−0.004]	(1.000)	0.076	[−0.010]	(1.000)	0.026	[−0.005]	(1.000)	0.046	[−0.013]	(1.000)
skew	−0.013	[−0.010]	(0.029)	−0.015	[−0.013]	(0.036)	**0.006**	[−0.002]	(0.463)	−**0.001**	[−0.005]	(0.313)
psl	0.011	[−0.002]	(0.945)	0.016	[−0.005]	(0.970)	0.010	[−0.002]	(0.937)	0.007	[−0.005]	(0.771)
roe	0.058	[−0.006]	(0.987)	0.074	[−0.010]	(0.967)	0.072	[−0.004]	(1.000)	0.080	[−0.011]	(1.000)
ia	0.020	[−0.012]	(0.967)	0.008	[−0.013]	(0.719)	0.038	[−0.004]	(0.975)	0.051	[−0.008]	(1.000)
qmj	0.052	[−0.007]	(0.976)	0.061	[−0.008]	(0.998)	0.128	[−0.004]	(0.982)	0.137	[−0.006]	(0.971)
bab	0.016	[−0.010]	(0.896)	−0.014	[−0.013]	(0.043)	0.045	[−0.003]	(0.989)	0.040	[−0.007]	(0.954)
gp	0.022	[−0.003]	(0.972)	0.020	[−0.009]	(0.951)	0.059	[−0.001]	(0.992)	0.055	[−0.006]	(0.984)
cma	0.001	[−0.009]	(0.341)	−0.009	[−0.012]	(0.137)	0.022	[−0.002]	(0.980)	0.023	[−0.005]	(0.967)
rmw	-0.009	[−0.019]	(0.147)	−0.016	[−0.020]	(0.086)	0.036	[−0.002]	(1.000)	0.043	[−0.006]	(0.992)
civ	0.014	[−0.009]	(0.981)	0.003	[−0.019]	(0.615)	0.015	[−0.008]	(0.991)	0.016	[−0.015]	(0.981)
		multiple test			multiple test			multiple test			multiple test	
	min	[−**0.022**]	**(0.122)**		[−**0.027**]	**(0.018)**		[−**0.011**]	**(0.997)**		[−**0.017**]	**(0.932)**

factor alone, the single-test *p*-value for the market factor is 0.001.

From a multiple-test perspective, we bootstrap to obtain the empirical distribution of the minimum statistic. In particular, following the bootstrap procedure in Section 2, we resample the time periods. For each bootstrapped sample, we first obtain the test statistic for each of the 14 orthogonalized factors and then record the minimum test statistic across all 14 statistics. The minimum statistic is the largest intercept reduction among the 14 factors. Since all factors are orthogonalized and therefore have no impact on the cross-section of expected returns, the minimum statistic shows what the largest intercept reduction can occur just by chance and therefore controls for multiple testing. It is important that all 14 test statistics are based on the same bootstrapped sample as this controls for test correlations, as emphasized by Fama and French (2010). Lastly, we compare the realized minimum statistic with the bootstrapped distribution of the minimum statistic to provide *p*-values.

Panel A of Table 2 shows the multiple testing results. In particular, the bootstrapped 5th percentile of SI^{med}_{ew} for the minimum statistic is −14.7%. By evaluating the 20.6% reduction against the empirical distribution of the minimum statistic for multiple testing, the *p*-value is 0.002. Therefore, the multiple-test *p*-value is below the 5% cutoff. We therefore also declare the market factor significant from a multiple testing perspective.

The fact that we always declare the market factor significant in our framework is not a trivial empirical find-

Table 3
Individual stocks as test assets, value weighted scaled intercepts.

This table shows test results for 14 risk factors using value-weighted individual stocks. (See Table I for the definitions of risk factors.) A stock needs to have at least 36 monthly observations (either in the original or the bootstrapped sample) to enter our tests. The baseline model refers to the model that includes the pre-selected risk factors. We focus on the panel regression model described in Section 2.1. The metric (i.e., SI^{m}_{vw}), which measures the difference in the value-weighted scaled absolute regression intercept, is defined in Section 3.3.2. Bold numbers are associated with the best incremental factor. The 5th percentile and *p*-value for multiple test correspond to the multiple-testing adjusted 5th percentile and the *p*-value for the best incremental factor, respectively.

Panel A: Baseline = No factor				Panel B: Baseline = *mkt*			Panel C: Baseline = *mkt+qmj*		
		single test			single test			single test	
Factor	SI_{vw}	5th-percentile	*p*-value	SI_{vw}	5th-percentile	*p*-value	SI_{vw}	5th-percentile	*p*-value
mkt	**−0.444**	[−0.258]	(0.000)						
smb	−0.059	[−0.054]	(0.041)	0.018	[−0.042]	(0.831)	0.076	[−0.032]	(0.994)
hml	0.144	[−0.059]	(0.972)	−0.038	[−0.045]	(0.128)	−0.016	[−0.062]	(0.471)
mom	0.153	[−0.064]	(1.000)	0.130	[−0.012]	(1.000)	0.125	[−0.026]	(1.000)
skew	−0.027	[−0.052]	(0.158)	−0.044	[−0.033]	(0.029)	−0.020	[−0.025]	(0.088)
psl	0.035	[−0.023]	(0.970)	0.016	[−0.011]	(0.996)	0.034	[−0.028]	(0.991)
roe	0.105	[−0.043]	(0.993)	−0.079	[−0.043]	(0.021)	0.038	[−0.025]	(0.967)
ia	0.382	[−0.086]	(0.984)	−0.042	[−0.048]	[0.083]	0.078	[−0.047]	(0.935)
qmj	0.363	[−0.112]	(0.892)	**−0.149**	[−0.079]	(0.002)			
bab	−0.048	[−0.035]	(0.026)	−0.088	[−0.049]	(0.006)	**−0.026**	[−0.037]	(0.157)
gp	−0.082	[−0.038]	(0.009)	−0.037	[−0.043]	(0.073)	−0.022	[−0.046]	(0.242)
cma	0.314	[−0.107]	(0.982)	−0.052	[−0.034]	(0.028)	0.019	[−0.038]	(0.941)
rmw	0.045	[−0.014]	(0.942)	−0.146	[−0.066]	(0.019)	0.053	[−0.033]	(1.000)
civ	−0.115	[−0.062]	(0.002)	0.035	[−0.019]	(0.973)	−0.017	[−0.024]	(0.113)
		Multiple test			Multiple test			Multiple test	
	min	**[−0.258]**	**(0.000)**		**[−0.083]**	**(0.004)**		**[−0.069]**	**(0.637)**

ing. Although the market factor has a strong theoretical motivation and is the first systematic risk factor tested (see Jensen et al., 1972), different papers, by using different testing methods and test assets, often arrive at conflicting conclusions. Therefore, there is no consensus as to whether the market factor is a valid risk factor empirically. Nonetheless, perhaps due to its intuitive appeal and theoretical relevance, most routinely use it for risk adjustment and cost of capital calculation. For instance, Fama and French (2015a) include the market factor in their five-factor model without testing its performance. More recently, by also using individual stocks as test assets and using the Fama–MacBeth cross-sectional test, (Jegadeesh et al., 2019) reject the market factor as a risk factor. Chordia et al. (2015) only find weak support for the market factor.

In contrast, our results suggest that the market factor is by far the dominant risk factor in explaining the cross-sectional variation in expected returns. We believe this is due to our use of the panel regression test.[24]

Fama (2015) summarizes the difference between the Fama–MacBeth cross-sectional approach and the panel regression approach. The panel regression approach essentially assumes that the factor risk premium is given by its in-sample estimate, and tries to evaluate what percentage of the expected excess return is explained by the risk exposure to the factor (i.e., beta times risk premium). The GRS test is one example of a panel regression test. It focuses on the extreme case in which the percentage of the expected excess return explained by the factor model has to be 100%, that is, the factor model is correctly specified and it is the underlying true factor model. Our test is less demanding than the GRS test in that this percentage does not have to be 100%. A factor could be declared useful as long as it explains a significant amount of expected returns for a given cross-section of assets. Indeed, in Tables 2 and 3, we show that the market factor single-handedly reduces the average pricing error by 44% and 21% with value weighting and equal weighting, respectively. These numbers are large from an economic perspective and much larger than the reduction in pricing error by alternative factors.

In the context of the long-standing debate on the role of the market factor, our results are consistent with Pukthuanthong et al. (2019), Kelly et al. (2019), and Giglio and Xiu (2019), while presenting a challenge to cross-sectional approaches as in Jegadeesh et al. (2019) and Chordia et al. (2015). Indeed, the choice of the testing framework (panel versus cross-sectional) is likely as important as (if not more than) adjusting for multiple tests and choosing a particular set of test assets, which are the other two concerns that our framework is able to address.

Why is there such a contrast between cross-sectional and time series (panel) methods? We believe that model misspecification is important. Specifically, time series (panel) regressions allow asset-specific intercepts to absorb idiosyncrasies that drive expected asset returns but are not captured by the regressors of interest (e.g., the market factor). As such, we are more likely to capture the incremental contribution of the regressors (in reducing the intercepts), thereby identifying the regressors as important. In contrast, cross-sectional regressions, by construction, only allow one intercept (across all assets) and are thus more susceptible

[24] Recent papers that support the market factor as a risk factor include Ahn et al., 2009; Bryzgalova, 2015; Cosemans et al., 2016; Berk and van Binsbergen, 2016; Bollerslev et al., 2016; Pukthuanthong et al., 2019; Hasler and Martineau, 2019a; 2019b, and Giglio and Xiu, 2019.

to model mis-specifications (i.e., omitted factors that drive the cross-section return difference but are not included in the regressor set). This is likely further exacerbated by the use of characteristic-sorted portfolios, which have a strong factor structure that is known to be unrelated to the market factor (Lewellen et al., 2010).

To better illustrate the intuition, we provide a simple example. Suppose assets A and B both have an excess return of 10%. A (B) has a market beta of one (two). Suppose the market risk premium is 5%. As such, A's excess return can be split into two 5% components, where one of them reflects the market risk premium multiplied by the beta and the other is due to omitted risk factors or could simply be mispricing. B's excess return is completely explained by CAPM. Under these assumptions, a cross-sectional regression would generate a zero slope coefficient; hence, CAPM is rejected. In contrast, our (panel regression) approach will correctly identify the market factor through the large reduction in intercept for both A and B, while leaving an unexplained alpha of 5% for A and a zero alpha for B.

In Panel B of Table 2, after the market factor is identified, *smb* is the best factor among the remaining factors. The percentage reduction in scaled absolute intercept is 6.2% under SI_{ew}^{med}. The corresponding multiple testing *p*-value is 0.039. We therefore declare *smb* significant. This result is perhaps not surprising given that equal weighting puts more weight on small stocks, which load heavily on *smb*. Nonetheless, the reduction in intercept associated with *smb* is second order compared to the market reduction.

In Panel C of Table 2, after both the market factor and *smb* are included in the baseline model, *hml* is the best factor among the remaining factors, reducing the scaled absolute intercept by 4.0% under SI_{ew}^{med}. It also has a significant multiple testing *p*-value under SI_{ew}^{med} (i.e., 1.8%). We therefore declare it significant and include it in the baseline model. After *mkt, smb*, and *hml* are included in the baseline model, none of the remaining factors is significant, as shown in Panel D of Table 2. We therefore terminate our testing procedure and identify the true factor model as *mkt+smb+hml* using the equally weighted test statistic.

In contrast to these results, which are consistent with Fama and French (1993), Fama and French (2015a) find highly significant alphas for *cma* and *rmw* in spanning tests that regress *cma* and *rmw* on existing factors. Why do they show up as insignificant in our framework? We want to stress that significant factors that survive spanning tests are not necessarily successful in explaining the returns of individual stocks. For example, Berkshire Hathaway's excess return, when viewed as a factor, generates an annualized alpha [against the Fama and French (2015a) five-factor model] of 10.6% with a *t*-statistic of 12.3. However, Berkshire Hathaway's excess return does not help explain the cross-section of stock returns and therefore would be declared useless in our tests.

In thinking about the economic magnitude of the pricing error reduction, the market factor reduces the baseline median scaled intercept by 20.6%. Conditional on the market factor, *smb*'s reduction of 6.2% implies a reduction of 4.9% ($= (1 - 20.6\%) \times 6.2\%$) of the baseline intercepts. Conditional on both the market factor and *smb, hml*'s reduction of 4.0% implies a reduction of 3.0% ($= (1 - 20.6\%) \times (1 - 6.2\%) \times 4.0\%$) of the baseline intercepts. As a result, the economic significance of *smb* and *hml* is modest compared to the market factor. The overall reduction (in terms of the baseline intercepts) is 28.5%, suggesting the limited success of the final model in explaining stock returns under equal weighting.

While our results with individuals stocks are consistent with Fama and French (1993), who propose a three-factor model based on size- and book-to-market-sorted portfolios, this does not mean that test assets do not matter for the testing outcomes. For example, as shown in our online appendix Section IB.2, we are unable to identify either *smb* or *hml* as useful risk factors using Fama–French 49-industry portfolios.

Contrary to our results, Chordia et al. (2015) and Jegadeesh et al. (2019) also use individual stocks and find that several popular factors (e.g., *smb* and *hml*) that are potentially risk factors do not seem to be priced. Both papers rely on the Fama–MacBeth regression (corrected for errors-in-variables bias) and use OLS in the second-stage regression. This method effectively equally weights the cross-section of stocks and is, therefore, consistent with our weighting scheme in Table 2. Our results thus show that the selected factor list also depends on the asset pricing test we use. Given that our method is less reliant on a fully specified model compared to alternative tests (as discussed previously), we believe that our approach could be more suitable for selecting risk factors based on a large cross-section of test assets.

Table 3 shows the results with value weighting of the pricing errors. Under value weighting, the market factor is again the best performing factor. Moreover, the economic magnitudes of the test statistics are much larger under value weighting than under equal weighting. For example, when the market factor is included in the baseline model, the reduction of the scaled absolute intercept is 44.4% under value weighting in Table 3, much greater than under equal weighting in Table 2.[25] Under value weighting, the multiple testing *p*-value for the market factor is less than 0.001, suggesting that the market factor is a highly significant risk factor. After the market factor is identified, the next best factor is *qmj*, which has a multiple testing *p*-value of 0.004. Note that the previously identified *hml* (under equal weighting) also implies a reduction of 3.8%. But this reduction is much smaller than the reduction of 14.9% by *qmj* under value weighting. As a result, *qmj* instead of *hml* is selected. After both the market factor and *qmj* are identified and included in the baseline model, none of the remaining factors is significant. Indeed, the multiple testing *p*-value for the next best factor (i.e., *bab*) is 0.637. Therefore, under value weighting, we find a two-factor model that includes *mkt* and *qmj*.

When we sequentially build the factor model, the drop in statistical significance for the best available candidate factor is remarkable. What is equally striking is the drop

[25] Our results are consistent with Plyakha et al. (2016), who show that individual stock alphas (relative to standard factor models) are higher under equal weighting than under value weighting.

in economic significance. For example, the market factor reduces the value-weighted absolute scaled intercept by 44.4%. After the market factor is included in the baseline model, the incremental reduction by the second identified factor (*qmj*) is 14.9% (which amounts to 8.3% ($= (1 - 44.4\%) \times 14.9\%$) in terms of the baseline intercepts). After both factors are included in the baseline model, the incremental reduction of the next best candidate (*bab*) is only 2.6% (i.e., 1.2% ($= (1 - 44.4\%) \times (1 - 14.9\%) \times 2.6\%$) in terms of the baseline intercepts). This drop in economic significance gives us confidence in the final model at which we arrive. Note that the baseline intercepts are reduced by 52.7% ($= 44.4\% + 8.3\%$) with the value weighted analysis, which roughly doubles the reduction using equal weighting (e.g., 28.5%).

Although our test picks up *qmj* as a useful risk factor, we want to stress that *qmj* is representative of a group of factors, that is, the *profitability group*. This group includes *qmj, rmw*, and *roe*. The three factors within the group are highly correlated and exhibit similar performance in our regression test.[26]

Our identification of a profitability factor under value weighting makes economic sense. Papers that propose profitability factors often use theories of firm investment to motivate their findings (e.g., Fama and French, 2015a; Hou et al., 2015). Intuitively, larger firms have fewer frictions and therefore can better engage in value maximization, the key assumption for investment theories to work. The fact that our test allows us to value weight the cross-section of intercepts demonstrates the flexibility of our approach.[27]

Our results using value weighting have important implications for the current practice of using portfolios as test assets in asset pricing tests. Average portfolio returns are disperse in the cross-section, which is helpful for asset pricing tests as dispersion potentially increases test power. However, for portfolios, the cross-section is small. Indeed, the dispersion of returns of the Fama–French 25 portfolios is largely driven by a few portfolios that are dominated by small stocks. Under equal weighting, current asset pricing tests are likely to identify factors that can explain these extreme portfolios. This is also consistent with it being relatively easy to data mine a factor that fits (by chance) these extreme portfolios.[28] However, this makes limited economic sense as portfolios that cover small stocks are less important than those that cover big stocks to an average investor whose portfolio is weighted in favor of big stocks.

Our results point to the market factor, conventional factors (i.e., *smb* and *hml*), and profitability factors. There are two additional aspects of the panel regression framework worth noting. First, we assume constant factor loadings while a Fama–MacBeth approach would have time-varying loadings. On the one hand, the reduction in estimation uncertainty in the panel framework for factor loadings could outweigh the increase in bias induced by fixed factor loadings. On the other hand, the constant factor loadings seem more appropriate for a portfolio approach in which stocks move in and out of characteristic-controlled portfolios. In the online appendix, Section IG.1, we show how to extend our framework to allow for time-varying factor loadings.

Second, estimating cross-sectional regressions as in the Fama–MacBeth approach is likely problematic for individual stocks as extreme observations in the cross-section are frequently observed. Trimming is an ill-advised practice as sometimes large observations provide important information for parameter estimates. In contrast, our panel regression framework focuses on the reduction in regression intercept or the *t*-statistic of intercept, both of which rely on the entire return time series and are less affected by a single observation.

3.3.4. Testing an augmented set of factors

We augment our factor list with 42 additional factors used by Ehsani and Linnainmaa (2021), totaling 56 factors. Our goal is to evaluate whether our results are sensitive to our choice of factor set.

Reporting all 56 factors in our tables is challenging. Therefore, on top of the 14 factors we studied before, we only list additional factors that generate a negative test statistic in our model (either SI_{ew}^{m} or SI_{ew}^{med}) once the market factor is included in our baseline model. The idea is that these factors have a higher likelihood of being selected to our final model. We do not find cases for factors for which our test statistics are positive (after the market factor is added to the baseline model), but the factors end up being selected to the final model. We explain the identities for these additional factors in the corresponding table legend.

Table 4 reports the results with equal-weighted test statistics and Table 5 with value-weighted test statistics. Our results in terms of the selection of the final factor model are the same as before: besides the market factor, size (*smb*) and book-to-market (*hml*) are found to be significant under equal weighting and quality-minus-junk (*qmj*, representative of the profitability group that also includes *rmw* and *roe*) is declared significant under value weighting. The primary difference from our results in Tables 2 and 3 lies in the multiple-testing adjusted test statistics. Due to a larger number of candidate factors, the best performing factor (i.e., the one that generates the minimum test statistic) needs to survive a tougher multiple testing threshold to claim significance. This tougher threshold is reflected in the fact that the multiple testing *p*-value is usually higher than the corresponding value in Tables 2 and 3. However, the decline in significance for the factors we select is usually small, highlighting the prominence of the 14 original factors on which we focus; their significance in pricing

[26] For example, *qmj* reduces the scaled absolute intercept by 14.9% and *rmw* by 14.6%. In our online appendix, Section IB.4, we conduct a limited experiment based on a comment made on the paper in which we combine factors that are highly correlated into new factors. In particular, we add two new factors (the average of *hml, ia*, and *cma* and the average of *qmj, roe*, and *rmw*) to the original 14 factors. Under either equal weighting or value weighting, combined factors do not seem to significantly improve the performance of the original factors.

[27] Chordia et al. (2015) use a modified Fama–MacBeth approach that corrects the bias in the return-premium estimation and find weak support for *rmw* as a priced risk factor. They do not consider *qmj*. The support for the profitability factors (both *rmw* and *qmj*) is much stronger in our model than in Chordia et al. (2015), which is likely due to both value weighting and our panel regression framework.

[28] See Lewellen et al. (2010) for a similar argument.

Table 4

Individual stocks as test assets, equally weighted scaled intercepts, 56 factors.

This table shows test results for 56 risk factors using equally weighted individual stocks. A stock needs to have at least 36 monthly observations (either in the original or the bootstrapped sample) to enter our tests. The baseline model refers to the model that includes the pre selected risk factors. We focus on the panel regression model described in Section 2.1. The two metrics (i.e., SI^{m}_{ew} and SI^{med}_{ew}), which measure the difference in equally weighted scaled mean/median absolute regression intercept, are defined in Section 3.2. For listed factors, the top 14 correspond to our focus group of factors defined in Table I. The rest correspond to factors that generate a negative test statistic in our model (either SI^{m}_{ew} or SI^{med}_{ew}) once the market factor is included in our baseline model. They are *cp* (cash flow to price), *di* (debt issuance), *ep* (earnings to price ratio), *em* (enterprise multiple), *lev* (leverage), *sp* (sales to price ratio), *ssi* (one-year share issuance), *lsi* (five-year share issuance), *maxr* (maximum daily return). Bold numbers are associated with the best incremental factor. The 5th percentile and *p*-value for multiple test corresponds to the multiple-testing adjusted 5th percentile and the *p*-value for the best incremental factor, respectively.

	Panel A: Baseline = No factor						Panel B: Baseline = *mkt*					
		Single test			Single test			Single test			single test	
Factor	SI^{m}_{ew}	5th-percentile	*p*-value	SI^{med}_{ew}	5th-percentile	*p*-value	SI^{m}_{ew}	5th-percentile	*p*-value	SI^{med}_{ew}	5th-percentile	*p*-value
mkt	−**0.192**	[−0.093]	(0.003)	−**0.206**	[−0.095]	(0.001)						
smb	−0.081	[−0.081]	(0.056)	−0.109	[−0.117]	(0.061)	−**0.041**	[−0.045]	(0.063)	−**0.062**	[−0.052]	(0.032)
hml	0.088	[−0.022]	(0.983)	0.108	[−0.029]	(1.000)	−0.021	[−0.030]	(0.131)	−0.047	[−0.028]	(0.014)
mom	0.091	[−0.034]	(1.000)	0.110	[−0.044]	(1.000)	0.070	[−0.007]	(1.000)	0.089	[−0.012]	(1.000)
skew	−0.008	[−0.031]	(0.278)	−0.002	[−0.034]	(0.478)	−0.004	[−0.009]	(0.167)	−0.003	[−0.013]	(0.319)
psl	0.011	[−0.019]	(0.920)	0.002	[−0.030]	(0.682)	0.001	[−0.004]	(0.409)	−0.003	[−0.012]	(0.237)
roe	0.163	[−0.042]	(0.951)	0.187	[−0.064]	(1.000)	0.142	[−0.019]	(1.000)	0.180	[−0.029]	(1.000)
ia	0.264	[−0.040]	(1.000)	0.291	[−0.048]	(1.000)	0.027	[−0.009]	(0.968)	0.015	[−0.015]	(0.934)
qmj	0.316	[−0.072]	(0.995)	0.358	[−0.090]	(0.998)	0.149	[−0.024]	(0.972)	0.193	[−0.029]	(0.973)
bab	−0.006	[−0.039]	(0.594)	−0.049	[−0.050]	(0.107)	0.018	[−0.010]	(0.983)	−0.014	[−0.017]	(0.181)
gp	0.017	[−0.008]	(0.529)	0.030	[−0.007]	(0.727)	0.023	[−0.005]	(0.961)	0.017	[−0.007]	(0.790)
cma	0.176	[−0.034]	(1.000)	0.199	[−0.035]	(1.000)	−0.012	[−0.013]	(0.057)	−0.031	[−0.019]	(0.027)
rmw	0.116	[−0.011]	(0.986)	0.137	[−0.017]	(0.994)	0.040	[−0.014]	(1.000)	0.048	[−0.020]	(0.975)
civ	−0.096	[−0.044]	(0.023)	−0.130	[−0.062]	(0.031)	−0.018	[−0.018]	(0.052)	−0.049	[−0.030]	(0.021)
cp	−0.023	[−0.009]	(0.023)	−0.024	[−0.011]	(0.021)	0.002	[−0.003]	(0.121)	−0.013	[−0.006]	(0.021)
di	0.009	[−0.007]	(0.821)	0.007	[−0.009]	(0.734)	0.003	[−0.001]	(0.306)	−0.001	[−0.003]	(0.119)
ep	−0.020	[−0.011]	(0.015)	−0.019	[−0.013]	(0.029)	0.000	[−0.000]	(0.084)	−0.004	[−0.003]	(0.038)
em	−0.023	[−0.008]	(0.018)	−0.028	[−0.013]	(0.009)	0.006	[−0.000]	(0.363)	−0.005	[−0.002]	(0.023)
lev	−0.002	[−0.009]	(0.200)	−0.001	[−0.010]	(0.407)	0.001	[−0.003]	(0.215)	−0.002	[−0.005]	(0.191)
sp	−0.008	[−0.011]	(0.081)	−0.011	[−0.014]	(0.068)	0.011	[−0.004]	(0.840)	−0.006	[−0.008]	(0.123)
ssi	−0.031	[−0.014]	(0.014)	−0.039	[−0.022]	(0.007)	0.002	[−0.004]	(0.422)	−0.011	[−0.012]	(0.052)
lsi	−0.020	[−0.019]	(0.053)	−0.026	[−0.025]	(0.043)	0.000	[−0.006]	(0.158)	−0.018	[−0.012]	(0.008)
maxr	−0.018	[−0.011]	(0.029)	−0.017	[−0.015]	(0.046)	−0.001	[−0.002]	(0.111)	0.001	[−0.006]	(0.391)
		Multiple test			Multiple test			Multiple test			multiple test	
	min	**[−0.123]**	**(0.000)**		**[−0.136]**	**(0.000)**		**[−0.045]**	**(0.071)**		**[−0.057]**	**(0.039)**

	Panel C: Baseline = *mkt*+*smb*						Panel D: Baseline = *mkt* + *smb*+*hml*					
		single test			single test			single test			single test	
Factor	SI^{m}_{ew}	5th-percentile	*p*-value	SI^{med}_{ew}	5th-percentile	*p*-value	SI^{m}_{ew}	5th-percentile	*p*-value	SI^{med}_{ew}	5th-percentile	*p*-value
mkt												
smb												
hml	−**0.017**	[−0.020]	(0.061)	−**0.040**	[−0.025]	(0.011)						
mom	0.055	[−0.004]	(1.000)	0.076	[−0.010]	(1.000)	0.026	[−0.005]	(1.000)	0.046	[−0.013]	(1.000)
skew	−0.013	[−0.010]	(0.029)	−0.015	[−0.013]	(0.036)	**0.006**	[−0.002]	(0.463)	−**0.001**	[−0.005]	(0.313)
psl	0.011	[−0.002]	(0.945)	0.016	[−0.005]	(0.970)	0.010	[−0.002]	(0.937)	0.007	[−0.005]	(0.771)
roe	0.058	[−0.006]	(0.987)	0.074	[−0.010]	(0.967)	0.072	[−0.004]	(1.000)	0.080	[−0.011]	(1.000)
ia	0.020	[−0.012]	(0.967)	0.008	[−0.013]	(0.719)	0.038	[−0.004]	(0.975)	0.051	[−0.008]	(1.000)
qmj	0.052	[−0.007]	(0.976)	0.061	[−0.008]	(0.998)	0.128	[−0.004]	(0.982)	0.137	[−0.006]	(0.971)
bab	0.016	[−0.010]	(0.896)	−0.014	[−0.013]	(0.043)	0.045	[−0.003]	(0.989)	0.040	[−0.007]	(0.954)
gp	0.022	[−0.003]	(0.972)	0.020	[−0.009]	(0.951)	0.059	[−0.001]	(0.992)	0.055	[−0.006]	(0.984)
cma	0.001	[−0.009]	(0.341)	−0.009	[−0.012]	(0.137)	0.022	[−0.002]	(0.980)	0.023	[−0.005]	(0.967)
rmw	−0.009	[−0.019]	(0.147)	−0.016	[−0.020]	(0.086)	0.036	[−0.002]	(1.000)	0.043	[−0.006]	(0.992)
civ	0.014	[−0.009]	(0.981)	0.003	[−0.019]	(0.615)	0.015	[−0.008]	(0.991)	0.016	[−0.015]	(0.981)
cp	0.014	[−0.002]	(0.772)	0.015	[−0.001]	(0.819)	0.015	[0.003]	(0.833)	0.017	[0.000]	(0.807)
di	0.003	[−0.000]	(0.227)	0.007	[−0.003]	(0.573)	0.003	[0.002]	(0.181)	0.007	[−0.002]	(0.523)
ep	0.005	[−0.001]	(0.360)	0.007	[−0.002]	(0.588)	0.008	[0.002]	(0.575)	0.013	[−0.001]	(0.800)
em	0.017	[−0.003]	(0.951)	0.015	[−0.001]	(0.904)	0.016	[0.003]	(0.940)	0.014	[−0.001]	(0.848)
lev	0.007	[−0.001]	(0.553)	0.010	[−0.003]	(0.843)	0.007	[0.002]	(0.582)	0.006	[−0.002]	(0.520)
sp	0.027	[−0.004]	(1.000)	0.025	[−0.001]	(1.000)	0.024	[0.002]	(1.000)	0.024	[−0.000]	(1.000)
ssi	0.012	[−0.000]	(0.902)	0.015	[−0.004]	(0.953)	0.018	[0.000]	(0.983)	0.024	[−0.002]	(0.997)
lsi	0.017	[−0.003]	(0.887)	0.016	[−0.000]	(0.877)	0.020	[0.002]	(0.945)	0.017	[0.002]	(0.842)
maxr	−0.000	[−0.002]	(0.113)	−0.000	[−0.007]	(0.314)	0.002	[0.000]	(0.187)	0.002	[−0.004]	(1.000)
		multiple test			multiple test			multiple test			multiple test	
	min	**[−0.021]**	**(0.181)**		**[−0.025]**	**(0.011)**		**[−0.011]**	**(0.981)**		**[−0.014[**	**(0.966)**

Table 5
Individual stocks as test assets, value weighted scaled intercepts, 56 factors.

This table shows test results for 56 risk factors using value-weighted individual stocks. A stock needs to have at least 36 monthly observations (either in the original or the bootstrapped sample) to enter our tests. The baseline model refers to the model that includes the pre selected risk factors. We focus on the panel regression model described in Section 2.1. The metric (i.e., SI_{vw}^{m}), which measures the difference in the value-weighted scaled absolute regression intercept, is defined in Section 3.3.2. For listed factors, the top 14 correspond to our focus group of factors defined in Table I. The rest correspond to factors that generate a negative test statistic in our model (either SI_{ew}^{m} or SI_{ew}^{med}) once the market factor is included in our baseline model. They are *ag* (asset growth), *noa* (net operating assets), *os* (*O*-score), *pm* (profit margin), *ic* (industry concentration), *sg* (sales growth), *ssi* (one-year share issuance), *lsi* (five-year share issuance), *sg* (sustainable growth), *tef* (total external financing), *mbeta* (market beta), *fage* (firm age), *ivol* (idiosyncratic volatility), *lr* (long-term reversals), *maxr* (maximum daily return). Bold numbers are associated with the best incremental factor. The 5th percentile and *p*-value for multiple test correspond to the multiple-testing adjusted 5th percentile and the *p*-value for the best incremental factor, respectively.

	Panel A: Baseline = No factor			Panel B: Baseline = *mkt*			Panel C: Baseline = *mkt*+*qmj*		
		Single test			Single test			Single test	
Factor	SI_{vw}	5th-percentile	*p*-value	SI_{vw}	5th-percentile	*p*-value	SI_{vw}	5th-percentile	*p*-value
mkt	−**0.444**	[−0.258]	(0.000)						
smb	−0.059	[−0.054]	(0.041)	0.018	[−0.042]	(0.831)	0.076	[−0.032]	(0.994)
hml	0.144	[−0.059]	(0.972)	−0.038	[−0.045]	(0.128)	−0.016	[−0.062]	(0.471)
mom	0.153	[−0.064]	(1.000)	0.130	[−0.012]	(1.000)	0.125	[−0.026]	(1.000)
skew	−0.027	[−0.052]	(0.158)	−0.044	[−0.033]	(0.029)	−0.020	[−0.025]	(0.088)
psl	0.035	[−0.023]	(0.970)	0.016	[−0.011]	(0.996)	0.034	[−0.028]	(0.991)
roe	0.105	[−0.043]	(0.993)	−0.079	[−0.043]	(0.021)	0.038	[−0.025]	(0.967)
ia	0.382	[−0.086]	(0.984)	−0.042	[−0.048]	[0.083]	0.078	[−0.047]	(0.935)
qmj	0.363	[−0.112]	(0.892)	−**0.149**	[−0.079]	(0.002)			
bab	−0.048	[−0.035]	(0.026)	−0.088	[−0.049]	(0.006)	−**0.026**	[−0.037]	(0.157)
gp	−0.082	[−0.038]	(0.009)	−0.037	[−0.043]	(0.073)	−0.022	[−0.046]	(0.242)
cma	0.314	[−0.107]	(0.982)	−0.052	[−0.034]	(0.028)	0.019	[−0.038]	(0.941)
rmw	0.045	[−0.014]	(0.942)	−0.146	[−0.066]	(0.019)	0.053	[−0.033]	(1.000)
civ	−0.115	[−0.062]	(0.002)	0.035	[−0.019]	(0.973)	−0.017	[−0.024]	(0.113)
ag	−0.050	[−0.022]	(0.000)	−0.008	[−0.003]	(0.013)	−0.007	[−0.003]	(0.891)
noa	−0.041	[−0.025]	(0.011)	−0.007	[−0.005]	(0.032)	0.011	[−0.007]	(0.968)
os	−0.001	[−0.019]	(0.432)	−0.002	[−0.010]	(0.429)	−0.001	[−0.010]	(0.470)
pm	−0.005	[−0.030]	(0.368)	−0.001	[−0.006]	(0.321)	−0.001	[−0.007]	(0.359)
ic	0.003	[−0.019]	(0.757)	−0.002	[−0.006]	(0.248)	0.002	[−0.006]	(0.650)
sg	−0.004	[−0.024]	(0.270)	−0.002	[−0.009]	(0.262)	−0.003	[−0.006]	(0.273)
ssi	−0.062	[−0.042]	(0.021)	−0.012	[−0.006]	(0.009)	0.008	[−0.005]	(0.962)
lsi	−0.047	[−0.033]	(0.033)	−0.007	[−0.005]	(0.031)	−0.004	[−0.006]	(0.121)
sg	−0.036	[−0.024]	(0.018)	−0.010	[−0.006]	(0.022)	0.004	[−0.002]	(0.698)
tef	−0.090	[−0.047]	(0.000)	−0.006	[−0.004]	(0.020)	0.003	[−0.003]	(0.792)
mbeta	−0.004	[−0.028]	(0.401)	−0.001	[−0.006]	(0.351)	−0.000	[−0.005]	(0.292)
fage	−0.021	[−0.033]	(0.068)	−0.005	[−0.004]	(0.028)	0.004	[−0.003]	(0.791)
ivol	−0.026	[−0.040]	(0.132)	−0.002	[−0.005]	(0.223)	−0.003	[−0.006]	(0.130)
lr	−0.019	[−0.022]	(0.091)	−0.005	[−0.004]	(0.041)	0.002	[−0.003]	(0.407)
maxr	−0.035	[−0.041]	(0.069)	−0.002	[−0.005]	(0.237)	−0.007	[−0.006]	(0.033)
		Multiple test			Multiple test			Multiple test	
	min	**[−0.213]**	**(0.000)**		**[−0.073]**	**(0.002)**		**[−0.070]**	**(0.689)**

stock returns cannot be crowded out by the additional 42 factors.

Overall, our results appear robust to the set of candidate factors we include in our tests. While we cannot exclude the existence of alternative factors given the large number of factors studied in the literature, we are able to identify a small subset in our tests.

3.3.5. Sorted portfolios

To contrast with our main results based on individual stocks, we apply our approach to a variety of sorted portfolios, including the Fama–French 25 portfolios, beta-sorted portfolios, Fama–French 49-industry portfolios, and 90 low-turnover anomaly portfolios as studied in Novy-Marx and Velikov (2016) and Kozak et al. (2018). We highlight the instability of the identified factor set across portfolios. We also discuss the issues associated with the use of the GRS statistic. We provide details of our results in online appendix, Section IF.

3.4. Robustness

A number of robustness tests are presented in the online appendix, Section IB, including trimming the smallest 10% of stocks for the individual stock analysis, value-weighted metrics for the portfolio results, block bootstrapping, and controls for infrequent trading.

3.5. Time-varying factor loadings and other issues

We discuss several issues that are related to our approach, including time-varying factor loadings, stepwise model selection, spurious factors, and factor model uncertainty and model misspecification. We provide details in our online appendix, Section IG.

We further explore the issue of time-varying factor loadings. In our model, we essentially assume the stationarity of the return-generating process for each stock and use unconditional regressions to obtain the factor load-

Table 6
Individual stocks as test assets, A pooled regression approach.
This table shows estimates from a pooled panel regression with time-varying factor loadings. We estimate the pooled regression model in Eq. (11) with individual stocks (see Harvey and Liu, 2020b). Standard errors are clustered by year-month and Fama–French 49 industries.

	Estimate	[Robust t-stat]
$Intercept$	0.009	[15.99]
$size_{i,t}$	−0.011	[−18.73]
smb_t	1.187	[62.38]
$size_{i,t} \times smb_t$	−0.911	[−33.60]
$btm_{i,t}$	0.002	[2.83]
hml_t	−0.44	[−17.71]
$btm_{i,t} \times hml_t$	1.198	[33.82]
mkt_t	0.893	[58.09]
$size_{i,t} \times mkt_t$	−0.289	[−16.64]
$btm_{i,t} \times mkt_t$	−0.168	[−7.40]
R-square	0.180	
Adjusted R-square	0.180	

ings. However, when conditioning information is available, we should be able to enrich the return-generating process by modeling the conditional distribution of returns as dependent on conditioning information. We analyze such a model below.

In particular, we focus on three factors identified in our main analysis under equal weighting: the market factor, *smb*, and *hml*. To capture time-varying stock characteristics, we include a stock's time-varying characteristic ranking in the regression model, following Harvey and Liu (2020b). Moreover, to capture time-varying factor loadings, we also allow characteristic rankings to affect the sensitivity to factor returns. Our model is represented by:

$$\begin{aligned} R_{it} - R_{ft} = {} & a_0 + b_{size,1} \cdot size_{i,t-1} + b_{size,2} \cdot smb_t \\ & + b_{size,3} \cdot size_{i,t-1} \times smb_t + b_{btm,1} \cdot btm_{i,t-1} \\ & + b_{btm,2} \cdot hml_t + b_{btm,3} \cdot btm_{i,t-1} \times hml_t \\ & + b_{mkt,1} \cdot mkt_t + b_{mkt,size} \cdot size_{i,t-1} \times mkt_t \\ & + b_{mkt,btm} \cdot btm_{i,t-1} \times mkt_t + \varepsilon_{i,t}, \end{aligned} \tag{11}$$

where $size_{i,t-1}$ and $btm_{i,t-1}$ are the lagged size and book-to-market rankings, and mkt_t denotes market excess return. We use a pooled panel regression to estimate our model. We also use robust standard errors that cluster by time (i.e., year-month) and Fama–French 49 industries.

Our model is related to those in the literature (e.g., Ferson and Harvey, 1999 and Avramov and Chordia, 2006) that model the interaction between factor loadings and firm characteristics. Our model differs from this literature in that it imposes a constant loading (across all firm-month observations) on the interaction term between characteristic ranking and factor returns. Our simplification is motivated by the economic restriction that is implicitly assumed in Daniel et al. (1997) and Bessembinder, Cooper, and Zhang (2019): once a stock is matched to a certain characteristic-sorted portfolio, its benchmark return is simply the portfolio return—no further risk adjustment (by running time series regressions) is needed.

Table 6 reports our results. All variables are highly significant in our model, suggesting the power of using individual stocks to identify important variables that drive the cross-section of stock returns. Notably, the four interaction terms are also significant, providing strong evidence for the importance of time-varying factor loadings. In addition, characteristics themselves also seem to have incremental power in explaining returns relative to characteristic-sorted portfolios themselves (with time-varying factor loadings). In contrast, our framework in Section 3.3 focuses on unconditional returns and thus cannot account for time-varying factor loadings or characteristics. The evidence presented in Table 6, while supporting our main idea of using individual stocks to test risk factors, points to the use of pooled panel regressions that incorporate time-varying stock characteristics as a direction for future research. We refer readers to Harvey and Liu (2020b) for details of the implementation of such pooled regression models, as well as model selection results that answer the question: what characteristics and factors help explain the cross-section of individual stock returns?

3.6. Caveats in applying our framework

Readers who are interested in applying our approach need to exert caution in two respects. First, the particular testing framework (e.g., panel regression versus cross-sectional regression) could have a large impact on the testing outcome, regardless of the multiple-testing correction. For individual stocks as test portfolios, we advocate the use of panel regressions due to concerns over influential observations and likely mis-specified models.

Second, when testing the incremental contribution of a proposed factor, the list of candidate factors and pre-determined factors must be fixed ex ante. Manipulating the list of candidate factors constitutes another source of multiple testing bias. Once this list is determined, the sequential nature of our stepwise selection approach will not affect the final testing outcome because, since pre-determined factors are also fixed, only a one-shot decision (i.e., does the proposed factor stand out from the list of candidate factors) needs to be made.

Third, profitability shows up in our tests, but it was only discovered in 2015. Essentially, the entire history of profitability in our examination is a backtest. What does it mean that a factor is useful, when it was not discovered? The advantage of market, *smb*, and *hml* is that they have longer histories (at least 30 years).

4. Conclusions

The finance profession has been on a 50-year quest to identify factors that explain the cross-section of expected returns. However, even after all this time, there is no consensus as to what the factor structure looks like. Our paper does not find the Holy Grail. However, we do make some progress on a number of vexing questions.

First, why is it that the choice of estimation (panel time series methods vs. cross-sectional Fama–MacBeth methods) leads to different factors being identified? Indeed, there are other puzzling issues, like the obvious dominance of the market factor in explaining the variance of individual stock returns as well as portfolio returns, yet it struggles to explain the cross-section of expected returns. We argue that the panel estimation method has a number of advantages and is likely more resilient to model misspecifications such as omitted variables. Further, we demonstrate that the method does not suffer from a situation in which the market factor works in the time series but not in the cross-section. Our results show that the market return is the single most important factor in explaining the cross-section of expected returns.

Second, why is it that portfolios based on different sorting characteristics lead to different sets of identified factors? For example, the factors identified with size and book-to-market sorts might be different from those identified with industry-sorted portfolios. Our solution is straightforward. We avoid portfolios and operate on individual stocks. Importantly, our simulation study shows that our test power with individual stocks is on a par with or greater when using portfolios, debunking a common notion that individual stocks are too noisy to use as test assets.

Third, why are asset pricing tests so reliant on the classic GRS test? It is well known that this test has many limitations. It can only be applied to a small number of assets, it often implies unrealistic short positions, it almost always rejects, and it is routinely used (problematically) as a heuristic in model comparison. We emphasize that this test will identify “factors” that have large excess returns (like Berkshire Hathaway stock) but little ability to explain the cross-section of expected returns.

Fourth, what do we do about test multiplicity given that hundreds of factors have been proposed? We propose a stepwise model selection method based on a bootstrapping approach. While many studies have used bootstrapping for model selection (e.g., Romano and Wolf, 2005), we offer a unique implementation that allows us to identify incremental factors. Crucially, we conditionally demean/adjust candidate factors so they have zero explanatory power for the cross-section of expected returns but retain their time series properties (such as correlation and non-normalities). We then simulate under the null and compare the empirical distribution to the actual factor properties. While we apply our method to factor selection with panel regressions, the method is general and can be used for cross-sectional methods as well as predictive regressions.

While we do not purport to answer all of these questions, we record some measured progress. That said, there is much work to do. For example, should we treat two candidate factors differently if one has a 30-year history and the other has recently been discovered (and potentially data mined)? Recent research in Chen and Zimmermann (2021) and Jensen et al. (2021) highlights the replicability of many anomalies discovered in the past. But just because some factor reproduces in the in-sample data does not mean the factor is “true”. That is, it is much more convincing if the in-sample results are replicated in either post-publication or on alternative assets (e.g., non-US assets).

There are also a number of challenges that we do not address when testing with individual stocks. For example, individual stocks’ characteristics will vary through time, leading to instability in risk loadings, perhaps more so with a portfolio approach that allows stocks to move in and out of the portfolio as their characteristics change. These questions are also left for future research.

Appendix A. A simulation study

A full-blown simulation study that takes all aspects of our method into account (e.g., the error rate for the first factor to be falsely identified, the error rate for the second factor to be falsely identified conditional on the first factor being correctly identified, etc.) is beyond the scope of this paper.[29] Our main goal for this simulation study is to evaluate the power of our bootstrap-based test in correctly identifying a risk factor that has incremental contribution (relative to the market factor) in explaining the cross-section of individual stock returns. This is motivated by the fact that the market factor is always found to be the most significant factor in our empirical study.

We first focus on firms that have a complete return history for the past 20 years. This gives us a balanced panel with $N=$ 2,732 firms in the cross-section and $T=240$ months in time series. A balanced panel is not required for our method to work. However, we use a balanced panel in our simulation study as it allows us to fix the number of firms in the cross-section. This lets us better evaluate how the test power changes with the length of the return time series.

We assume that a two-factor model (i.e., the market factor plus a candidate factor denoted as f_t) is the true model. We construct the panel of returns corresponding to the true model by sampling from the real data. In particular, we first project stock returns onto the two factors:

$$R_{it} - R_{ft} = \alpha_i + \beta_{i,m} mkt_t + \beta_{i,f} f_t + \varepsilon_{i,t}.$$

Let $\mathbf{e}_i = [\varepsilon_{i,1}, \varepsilon_{i,2}, \ldots, \varepsilon_{i,T}]'$ denote the vector of factor model residuals for stock i. We collect the cross-section of factor loadings and residuals into matrices $\mathbf{B}$ and $\mathbf{E}$:

$$\mathbf{B}_{(2\times N)} = [[\beta_{1,m}, \beta_{2,m}, \ldots, \beta_{N,m}]', [\beta_{1,f}, \beta_{2,f}, \ldots, \beta_{N,f}]']',$$
$$\mathbf{E}_{(T\times N)} = [\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_T].$$

We also project the candidate factor f_t onto the market factor mkt_t:

$$f_t = \alpha_f + \beta_m mkt_t + \varepsilon_{f,t}. \qquad \text{(A1)}$$

The magnitude of α_f determines how “true” the candidate factor is after its correlation with the market factor is taken into account. For example, $\alpha_f = 0$ means that the risk premium of the candidate factor is explained by its exposure to the market factor, so it has a zero incremental contribution to explaining the cross-section of expected returns. This constitutes the null hypothesis. Table (A.1) summarizes α_f for all candidate factors for the past 20 years. For our follow-up analysis, we choose to present results for the top five factors based on the ranking of their t-statistics for α_f. Results for the other factors are similar.

To evaluate the test power corresponding to different alternative hypotheses regarding the candidate factor, we

[29] See Harvey et al. (2016) for a discussion on test power when there are multiple hypothesis tests.

Table A.1
Summary statistics on α_f, January 1993–December 2012.

We project a candidate factor f_t onto the market factor mkt_t through the regression $f_t = \alpha_f + \beta_m mkt_t + \varepsilon_{f,t}$. We report the level and the t-statistic of the regression intercept α_f corresponding to the 14 risk factors. (See Table 1 for the definitions of risk factors.)

	smb	*hml*	*mom*	*skew*	*psl*	*roe*	*ia*	*qmj*	*bab*	*gp*	*cma*	*rmw*	*civ*
Mean	0.014	0.047	0.083	0.033	0.078	0.072	0.054	0.078	0.112	0.056	0.056	0.061	0.067
t-stat	[0.52]	[1.86]	[2.08]	[1.63]	[2.48]	[3.60]	[3.40]	[4.45]	[3.63]	[3.13]	[3.56]	[3.11]	[2.28]

assume that the true candidate factor is

$$f_t^A = A \times \alpha_f + \beta_m mkt_t + \varepsilon_{f,t}. \tag{A2}$$

By setting A at zero, the factor premium is completely explained by its exposure to the market factor. As a result, the candidate factor has zero incremental explanatory power for the cross-section of expected returns. This constitutes our null hypothesis. The test power corresponding to the null hypothesis tells us the size of the test. By setting A at other values, the alternative hypothesis is true. By changing the magnitude of A, we are able to evaluate the test power corresponding to different levels of factor premiums, which indicate how significant the candidate factor is in offering incremental information to explain the cross-section of expected returns.

Using the factor loadings and return residuals stored in **B** and **E**, we create the panel of returns corresponding to the true model. In particular, for a resampled time index $\{t_j^w\}_{j=1}^T$ and for a given level of A, the panel of excess returns is given by

$$rx_{i,j}^w = \beta_{i,m} mkt_{t_j^w} + \beta_{i,f} f_{t_j^w}^A + \varepsilon_{i,t_j^w} \tag{A3}$$

$$= \beta_{i,m} mkt_{t_j^w} + \beta_{i,f}(A \times \alpha_f + \beta_m mkt_{t_j^w} + \varepsilon_{f,t_j^w}) + \varepsilon_{i,t_j^w},\ j = 1, \ldots, T;\ \ i = 1, \ldots, N. \tag{A4}$$

The way we construct the return panel is slightly different from standard simulation methods in that, instead of using Gaussian variables to simulate return residuals, we use bootstrapped residuals based on the real data. This allows us to take the non normality in returns into account and, at the same time, maintain the dependency among the cross-section of realized return residuals, as emphasized by Fama and French (2010). Note that our way of using bootstrapping to construct the return panel implicitly assumes that time series observations are independent. While we do not separately study the impact of time series dependence on our tests in the simulation study, we provide robustness checks for our main results, which take time series dependence into account by performing block bootstrapping; see online appendix Section IA.1.

Let the simulated return panel corresponding to the wth resampled time index be $\mathbf{RX}^w$. For this sample, we use our method to make a decision as to whether the candidate factor is significant. Let $D_w = 1$ denote the event that the candidate factor is declared significant ($D_w = 0$ denotes otherwise). We bootstrap the time index W (= 1,000) times and use $\sum_{w=1}^W D_w/W$ to approximate the test power.

To compare our approach with alternative testing methods, we consider two popular methods based on beta sorts. The first method is unconditional beta sorts, which first sorts stocks based on their unconditional factor loadings estimated over the entire sample, and then forms the long-short portfolio by having a long position in stocks that are in the top beta decile and a short position in stocks that are in the bottom beta decile. The t-statistic of the portfolio's returns is used to test the significance of the candidate factor. The second method is conditional beta sorts. We estimate factor loadings based on a five-year rolling window and construct the long-short portfolio that we hold out of sample for one year. We again use the t-statistic of the portfolio returns to test the significance of the candidate factor.

To examine how the length of the time series affects test power, we double the length of the time series by creating a new return panel that fixes the cross-section and repeats the time series of the original panel. Essentially, we are assuming that returns are stationary so we can draw their future realizations from their past realizations. Similarly, we create new factor time series. We then follow the aforementioned procedures of the simulation study to examine test power when the sample size of the time series doubles.

Tables A.2 and A.3 show the simulation results for $T = 240$ and $T = 480$, respectively. When $T = 240$ and $A = 0$, the significance levels of all tests seem to be controlled at approximately 5%, which is the pre-specified significance level. However, the two beta sorts seem to under-reject the null as the Type I errors of many of the beta sorts are below 5%. This is likely because the size of the time series is small. When T is increased to 480, the significance levels of many beta sorts are higher and are closer to 5%.

When $A = 1.0$ (that is, the factor risk premium is the same as the original factor), the power of our tests based on the t-statistics is in general higher than that based on both types of beta sorts. In particular, when $T = 240$, the gain in power by using our tests is about 10% on average. However, the gain is not uniform across factors. For example, the power of our tests is similar to that based on the unconditional beta sorts for *qmj*, and is about 20% higher than both conditional and unconditional beta sorts for *cma*.

When $T = 480$, which is closer to the size of the 1968–2012 period that we examine in the paper, the power of our tests seems high. For $A = 1.0$ and for tests based on t-statistics, it ranges from 72% (*bab*) to 95% (*qmj*). For our application with the real data, we have $T = 540$. However, we do not have $N = 2{,}732$ firms that exist throughout the entire sample. The total number of firm-month observations in our sample is about 1.8 times that of the simulation study.[30] We therefore believe that our tests should have a high power for the real applications.

[30] We have slightly more than 20,000 firms in the cross-section. On average, a firm exists for around ten years. Therefore, our sample size is about 1.8 (= (20,000× 120)/(2,732 × 480)) times that of the simulation study.

Table A.2

Test power, individual stocks, $T = 240$.

This table shows test power for risk factors. For a given risk factor f, we project it onto the market return (Eq. (A.1)) to obtain the regression intercept. We then construct a new factor (f^A) based on Eq. (A.2), with A controlling the factor risk premium that is not explained by the market factor. We perform $D =$ 1,000 sets of simulations. For each set, we bootstrap the sample period to construct the time series for both the new factor given in Eq. (A.2) and the market factor, and, by assuming the two-factor model is true, construct a panel of returns through Eq. (A.4), where the return innovations are resampled from the return innovations based on the real data. Based on the bootstrapped factors and return panels, we test whether f^A has incremental power to explain the cross-section of expected returns. We calculate the test power by averaging the number of rejections across simulations. We consider four tests. Two of them, SI^{m}_{ew} and SI^{med}_{ew}, are explained in Section 3.2. The rest are two beta sorts. The unconditional beta sorts ($Uncond.\beta$) first sorts stocks based on their unconditional factor loadings estimated over the entire sample, and then forms the long-short portfolio by having a long position in stocks that are in the top beta decile and a short position in stocks that are in the bottom beta decile. The t-statistic of the portfolio's returns is used to test the significance of the candidate factor. The conditional beta sorts ($Cond.\beta$) first estimates factor loadings based on a five-year rolling window and then constructs the long-short portfolio that we hold out of sample for one year. The t-statistic of the portfolio returns is used to test the significance of the candidate factor. The five factors examined are explained in Table A.1. We set $T = 240$ and focus on a cross-section of 2,732 firms that have a complete return history for the past 20 years.

Factor	Method	$A = 0$ (null)	$A = 0.5$	$A = 1.0$	$A = 1.5$	$A = 2.0$
roe	SI^{m}_{ew}	0.022	0.167	0.517	0.729	0.865
	SI^{med}_{ew}	0.037	0.190	0.441	0.677	0.832
	$Cond.\beta$	0.016	0.113	0.377	0.663	0.836
	$Uncond.\beta$	0.008	0.055	0.309	0.709	0.885
ia	SI^{m}_{ew}	0.061	0.214	0.633	0.839	0.969
	SI^{med}_{ew}	0.067	0.213	0.597	0.854	0.947
	$Cond.\beta$	0.025	0.191	0.390	0.672	0.869
	$Uncond.\beta$	0.014	0.137	0.541	0.833	0.961
qmj	SI^{m}_{ew}	0.028	0.320	0.671	0.867	0.949
	SI^{med}_{ew}	0.039	0.278	0.608	0.812	0.896
	$Cond.\beta$	0.010	0.212	0.576	0.801	0.931
	$Uncond.\beta$	0.012	0.133	0.552	0.885	0.971
bab	SI^{m}_{ew}	0.047	0.248	0.600	0.809	0.947
	SI^{med}_{ew}	0.059	0.229	0.587	0.794	0.939
	$Cond.\beta$	0.031	0.134	0.375	0.642	0.828
	$Uncond.\beta$	0.023	0.197	0.463	0.773	0.939
cma	SI^{m}_{ew}	0.053	0.188	0.534	0.723	0.809
	SI^{med}_{ew}	0.043	0.208	0.469	0.714	0.817
	$Cond.\beta$	0.036	0.113	0.321	0.648	0.832
	$Uncond.\beta$	0.047	0.052	0.337	0.743	0.902

Table A.3

Test power, individual stocks, $T = 480$.

This table shows size-adjusted test power for risk factors. For a given risk factor f, we project it onto the market return (Eq. (A.1)) to obtain the regression intercept. We then construct a new factor (f^A) based on Eq. (A.2), with A controlling the factor risk premium that is not explained by the market factor. We perform $D =$ 1,000 sets of simulations. For each set, we bootstrap the sample period to construct the time series for both the new factor given in Eq. (A.2) and the market factor, and, by assuming the two-factor model is true, construct a panel of returns through Eq. (A.4) in which the return innovations are resampled from the return innovations based on the real data. Based on the bootstrapped factors and return panels, we test whether f^A has incremental power to explain the cross-section of expected returns. We calculate the test power by averaging the number of rejections across simulations. We consider four tests. Two of them, SI^{m}_{ew} and SI^{med}_{ew}, are explained in Section 3.2. The rest are two beta sorts. The unconditional beta sorts ($Uncond.\beta$) first sorts stocks based on their unconditional factor loadings estimated over the entire sample, and then forms the long-short portfolio by having a long position in stocks that are in the top beta decile and a short position in stocks that are in the bottom beta decile. The t-statistic of the portfolio's returns is used to test the significance of the candidate factor. The conditional beta sorts ($Cond.\beta$) first estimates factor loadings based on a five-year rolling window and then constructs the long-short portfolio that we hold out of sample for one year. The t-statistic of the portfolio returns is used to test the significance of the candidate factor. The five factors examined are explained in Table A.1. We set $T = 480$ by repeating the returns for a cross-section of 2,732 firms that have a complete return history for the past 20 years.

Factor	Method	$A = 0$ (null)	$A = 0.5$	$A = 1.0$	$A = 1.5$	$A = 2.0$
roe	SI^{m}_{ew}	0.024	0.277	0.801	0.937	0.985
	SI^{med}_{ew}	0.021	0.258	0.721	0.925	0.993
	$Cond.\beta$	0.013	0.179	0.536	0.843	0.973
	$Uncond.\beta$	0.020	0.093	0.525	0.895	0.992
ia	SI^{m}_{ew}	0.025	0.378	0.861	0.974	1.000
	SI^{med}_{ew}	0.016	0.351	0.834	0.973	1.000
	$Cond.\beta$	0.026	0.274	0.756	0.948	0.990
	$Uncond.\beta$	0.030	0.383	0.849	0.970	0.992
qmj	SI^{m}_{ew}	0.032	0.552	0.952	0.999	1.000
	SI^{med}_{ew}	0.027	0.486	0.946	1.000	1.000
	$Cond.\beta$	0.011	0.328	0.825	0.987	1.000
	$Uncond.\beta$	0.016	0.286	0.879	0.996	1.000
bab	SI^{m}_{ew}	0.028	0.289	0.737	0.941	0.998
	SI^{med}_{ew}	0.021	0.247	0.683	0.953	0.993
	$Cond.\beta$	0.015	0.155	0.591	0.894	0.991
	$Uncond.\beta$	0.008	0.127	0.603	0.935	0.997
cma	SI^{m}_{ew}	0.041	0.373	0.817	0.971	1.000
	SI^{med}_{ew}	0.037	0.312	0.788	0.978	1.000
	$Cond.\beta$	0.056	0.201	0.612	0.913	0.981
	$Uncond.\beta$	0.051	0.273	0.749	0.967	1.000

Table A.4
Test power adjusted for test size, individual stocks, $T = 480$.

This table shows test power for risk factors. For a given risk factor f, we project it onto the market return (Eq. (A.1)) to obtain the regression intercept. We then construct a new factor (f^A) based on Eq. (A.2), with A controlling the factor risk premium that is not explained by the market factor. We perform $D =$ 1,000 sets of simulations. For each set, we bootstrap the sample period to construct the time series for both the new factor given in Eq. (A.2) and the market factor, and, by assuming the two-factor model is true, construct a panel of returns through Eq. (A.4) in which the return innovations are resampled from the return innovations based on the real data. Based on the bootstrapped factors and return panels, we test whether f^A has incremental power to explain the cross-section of expected returns. We calculate the test power by averaging the number of rejections across simulations. To adjust test power by test size, we adjust the t-statistic cutoff under the null hypothesis such that the adjusted size exactly equals the pre specified nominal size (i.e., 5%). We then calculate the adjusted test power by averaging the number of rejections based on the adjusted t-statistic cutoff. We consider four tests. Two of them are SI^{m}_{ew} and SI^{med}_{ew} that are explained in Section 3.2. The rest are two beta sorts. The unconditional beta sorts ($Uncond.\beta$) first sorts stocks based on their unconditional factor loadings estimated over the entire sample, and then forms the long-short portfolio by having a long position in stocks that are in the top beta decile and a short position in stocks that are in the bottom beta decile. The t-statistic of the portfolios returns is used to test the significance of the candidate factor. The conditional beta sorts ($Cond.\beta$) first estimates factor loadings based on a five-year rolling window and then constructs the long-short portfolio that we hold out of sample for one year. The t-statistic of the portfolio returns is used to test the significance of the candidate factor. The five factors examined are explained in Table A.1. We set $T = 480$ by repeating the returns for a cross-section of 2,732 firms that have a complete return history for the past 20 years.

Factor	Method	$A = 0$ (null)	$A = 0.5$	$A = 1.0$	$A = 1.5$	$A = 2.0$
roe	SI^{m}_{ew}	0.050	0.309	0.904	1.000	1.000
	SI^{med}_{ew}	0.050	0.297	0.830	1.000	1.000
	$Cond.\beta$	0.050	0.214	0.642	0.998	1.000
	$Uncond.\beta$	0.050	0.106	0.598	1.000	1.000
ia	SI^{m}_{ew}	0.050	0.419	0.984	1.000	1.000
	SI^{med}_{ew}	0.050	0.425	0.969	1.000	1.000
	$Cond.\beta$	0.050	0.300	0.851	1.000	1.000
	$Uncond.\beta$	0.050	0.419	0.937	0.999	1.000
qmj	SI^{m}_{ew}	0.050	0.601	0.998	1.000	1.000
	SI^{med}_{ew}	0.050	0.542	1.000	1.000	1.000
	$Cond.\beta$	0.050	0.388	0.993	1.000	1.000
	$Uncond.\beta$	0.050	0.334	1.000	1.000	1.000
bab	SI^{m}_{ew}	0.050	0.329	0.812	0.978	1.000
	SI^{med}_{ew}	0.050	0.279	0.793	1.000	1.000
	$Cond.\beta$	0.050	0.185	0.680	0.983	1.000
	$Uncond.\beta$	0.050	0.154	0.740	0.991	1.000
cma	SI^{m}_{ew}	0.050	0.387	0.847	0.987	1.000
	SI^{med}_{ew}	0.050	0.328	0.833	0.990	1.000
	$Cond.\beta$	0.050	0.194	0.590	0.886	0.953
	$Uncond.\beta$	0.050	0.272	0.745	0.962	0.995

Our analysis above follows the standard approach in statistics that compares performance among alternative tests. In particular, we compare the powers of competing tests that have the same nominal size. However, differences in test power could simply reflect differences in test size. To address this concern, we report size-adjusted test power in Table A.4 (e.g., Harvey and Liu, 2020a). More specifically, for a given test reported in Table A.3, we adjust the t-statistic cutoff to a certain level such that the simulated test size exactly equals the pre-specified nominal size. We then calculate the simulated test power based on the adjusted significance level.

Compared to Table A.3, the difference in adjusted test power in Table A.4 is smaller between our approach and the two beta sorts. For example, in Table A.4, test power for *qmj* (with $A = 1.0$) is 99.8% for our approach with SI^{m}_{ew} and is 99.3% for conditional beta sorts. This difference is smaller than 12.7% ($= 95.2\% - 82.5\%$), as reported in Table A.3. This can be explained by the fact that given the smaller actual sizes for the two beta sorts in Table A.3, their test powers are increased more after size adjustment, leading to a smaller difference in test power between our approach and the two beta sorts. Nevertheless, the test power for our approach still compares favorably with the two beta sorts in most cases.

Overall, our simulation results based on individual stocks suggest that our tests, in particular the tests based on t-statistics, have high test power, both in an absolute and relative sense. When the length of the time series is close to our applications, the simulation results show that the power of our tests is well above 70%. It also compares favorably with the tests that are based on either unconditional or conditional beta sorts.

We redo the same simulation exercise based on the Fama-French 25 portfolios, as shown in Table A.5. For $A = 1.0$, our tests based on t-statistics again dominate those based on beta sorts. The increase in power of our tests SI^{med}_{ew} relative to the unconditional beta sorts (i.e., the better one between the two beta sorts) is again nonuniform across factors but is on average about 15%.

More interestingly, comparing Table A.5 with A.3, which has a similar number of time periods to Table A.5, we are not necessarily losing power by considering individual stocks. Focusing on our tests based on SI^{med}_{ew} and when $A = 1.0$, our tests based on individual stocks have a higher power for *roe* and *qmj* than those based on the Fama-French 25 portfolios. Overall, across the five factors, our tests based on individual stocks have a similar power to our tests based on the Fama-French 25 portfolios. This seems to be at odds with the conventional thinking that

Table A.5

Test power, Fama-French 25 portfolios, 1968–2012.

This table shows test power for risk factors. For a given risk factor f, we project it onto the market return (Eq. (A.1)) to obtain the regression intercept. We then construct a new factor (f^A) based on Eq. (A.2), with A controlling the factor risk premium that is not explained by the market factor. We perform $D =$ 1,000 sets of simulations. For each set, we bootstrap the sample period to construct the time series for both the new factor given in Eq. (A.2) and the market factor, and, by assuming the two-factor model is true, construct a panel of returns through Eq. (A.4) in which the return innovations are resampled from the return innovations based on the real data. Based on the bootstrapped factors and return panels, we test whether f^A has incremental power to explain the cross-section of expected returns. We calculate the test power by averaging the number of rejections across simulations. We consider four tests. Two of them, SI^m_{ew} and SI^{med}_{ew}, are explained in Section 3.2. The rest are two beta sorts. The unconditional beta sorts ($Uncond.\beta$) first sorts stocks based on their unconditional factor loadings estimated over the entire sample, and then forms the long-short portfolio by having a long position in stocks that are in the top beta decile and a short position in stocks that are in the bottom beta decile. The t-statistic of the portfolio's returns is used to test the significance of the candidate factor. The conditional beta sorts ($Cond.\beta$) first estimates factor loadings based on a five-year rolling window and then constructs the long-short portfolio that we hold out of sample for one year. The t-statistic of the portfolio returns is used to test the significance of the candidate factor. The five factors examined are explained in Table 1. We focus on the Fama-French 25 portfolios that cover the period 1968–2012.

Factor	Method	$A = 0$ (null)	$A = 0.5$	$A = 1.0$	$A = 1.5$	$A = 2.0$
roe	SI^m_{ew}	0.047	0.519	0.793	0.915	0.975
	SI^{med}_{ew}	0.043	0.448	0.765	0.890	0.956
	$Cond.\beta$	0.028	0.179	0.509	0.817	0.962
	$Uncond.\beta$	0.019	0.156	0.591	0.903	0.991
ia	SI^m_{ew}	0.033	0.668	0.981	1.000	1.000
	SI^{med}_{ew}	0.035	0.569	0.952	1.000	1.000
	$Cond.\beta$	0.014	0.263	0.817	0.992	1.000
	$Uncond.\beta$	0.012	0.281	0.836	0.990	1.000
qmj	SI^m_{ew}	0.044	0.572	0.886	0.978	0.998
	SI^{med}_{ew}	0.041	0.539	0.850	0.962	0.993
	$Cond.\beta$	0.025	0.274	0.752	0.971	0.992
	$Uncond.\beta$	0.017	0.279	0.841	0.993	1.000
bab	SI^m_{ew}	0.039	0.437	0.797	0.945	0.989
	SI^{med}_{ew}	0.035	0.357	0.748	0.919	0.980
	$Cond.\beta$	0.013	0.091	0.298	0.636	0.853
	$Uncond.\beta$	0.009	0.088	0.376	0.712	0.901
cma	SI^m_{ew}	0.049	0.542	0.946	0.999	1.000
	SI^{med}_{ew}	0.045	0.469	0.906	0.992	1.000
	$Cond.\beta$	0.019	0.205	0.689	0.958	1.000
	$Uncond.\beta$	0.039	0.257	0.766	0.971	0.999

individual stocks are more noisy and thus less informative than portfolios in factor tests. Our t-statistics-based tests, by taking the return volatility into account, seem to be able to detect a true factor as often as tests based on portfolios.

There are several takeaways from our simulation study. First, our tests based on t-statistics seem to be more powerful than those based on alternative statistics. We therefore favor our tests that are based on t-statistics in our applications.

Second, compared to traditional beta sorts, we are not losing power by using our bootstrap-based approach. In fact, we see a mild increase in power across a variety of factors by using both individual stocks and Fama-French 25 portfolios. While bootstrapping is not essential for the two-factor exercise we perform in the simulation study, it is key to our tests in the paper that build on the maximum/minimum test statistics. When extreme test statistics are used to adjust for multiple testing, traditional tests (e.g., conditional or unconditional beta sorts) are no longer appropriate as the asymptotic distributions for the extreme test statistics are not known in closed form. Moreover, we do not know how well the asymptotic distributions work in finite samples. Bootstrapping offers a convenient way to provide inference, as shown in White (2000) and Romano and Wolf (2005). It is therefore important to show that the bootstrap-based approach has power in the context of application.

Third, we are not losing power by considering individual stocks. The average performance of our tests based on individual stocks is similar to that of our tests based on the Fama-French 25 portfolios. The key assumption for our simulation study is that a two-factor model is the true underlying factor model, either for individual stocks or the Fama-French 25 portfolios. In reality, asset returns could be determined by a more complicated model. Compared to the Fama-French 25 portfolios, which are constructed to maximize the exposure to two existing factors, individual stocks potentially can provide unbiased and significantly richer information to identify the true factor model.

References

Affleck-Graves, J., McDonald, B., 1990. Nonnormalities and tests of asset pricing theories. J. Financ. 44, 889–908.

Ahn, D., Conrad, J., Dittmar, R., 2009. Basis assets. Rev. Financ. Stud. 22, 5133–5174.

Ang, A., Liu, J., Schwarz, K., 2016. Using Stocks or Portfolios in tests of Factor Models. Columbia University. Unpublished working paper

Asness, C., Frazzini, A., Pedersen, L.H., 2019. Quality minus junk. Rev. Account. Stud. 24, 34–112.

Avramov, D., Chordia, T., 2006. Asset pricing models and financial market anomalies. Rev. Financ. Stud. 19, 1001–1040.

Barillas, F., Shanken, J. A., 2017. Which alpha? Rev. Financ. Stud. 30,1316–1338.

Berk, J.B., 2000. Sorting out sorts. J. Financ. 55, 407–427.

Berk, J.B., van Binsbergen, J.H., 2016. Assessing asset pricing models using revealed preferences. J. Financ. Econ. 119, 1–23.

Bessembinder, H., Cooper, M.J., Zhang, F., 2019. Characteristic-based benchmark returns and corporate events. Rev. Financ. Stud. 32, 75–125.

Bollerslev, T., Li, S.Z., Todorov, V., 2016. Roughing up beta: continuous versus discontinuous betas and the cross section of expected stock returns. J. Financ. Econ. 120, 464–490.

Bryzgalova, S., 2015. Spurious factors in linear asset pricing models. LSE Unpublished working paper.

Carhart, M.M., 1997. On persistence in mutual fund performance. J. Financ. 52, 57–82.

Chen, A.Y., Zimmermann, T., 2021. Open Source Cross-Sectional Asset Pricing. Federal Reserve Board. Unpublished working paper

Chen, H., Chen, S., Chen, Z., Li, F., 2019. Empirical investigation of an equity pairs trading strategy. Manag. Sci. 65, 370–389.

Chen, L., Novy-Marx, R., Zhang, L., 2010. An Alternative Three-Factor Model. University of Rochester. Unpublished working paper

Chordia, T., Goyal, A., Shanken, J., 2015. Cross-Sectional Asset Pricing with Individual Stocks: Betas Versus Characteristics. Emory University. Unpublished working paper

Cosemans, M., Frehen, R., Schotman, P.C., Bauer, R., 2016. Estimating security betas using prior information based on firm fundamentals. Rev. Financ. Stud. 29, 1072–1112.

Croce, M.M., Marchuk, T., Schlag, C., 2019. The Leading Premium. Bocconi University. Unpublished working paper

Daniel, K., Grinblatt, M., Titman, S., Wermers, R., 1997. Measuring mutual fund performance with characteristic-based benchmarks. J. Financ. 52, 1035–1058.

Ehsani, S., Linnainmaa, J.T., 2021. Factor momentum and the momentum factor. J. Financ.. forthcoming

Fama, E.F., 1998. Market efficiency, long-term returns, and behavioral finance. J. Financ. Econ. 42, 283–306.

Fama, E.F., 2015. Cross-section Versus Time-Series Tests of Asset Pricing Models. University of Chicago. Unpublished working paper

Fama, E.F., French, K.R., 1993. Common risk factors in the returns on stocks and bonds. J. Financ. Econ. 33, 3–56.

Fama, E.F., French, K.R., 2010. Luck versus skill in the cross-section of mutual fund returns. J. Financ. 65, 1915–1947.

Fama, E.F., French, K.R., 2015a. A five-factor asset pricing model. J. Financ. Econ. 116, 1–22.

Fama, E.F., French, K.R., 2015b. Incremental variables and the investment opportunity set. J. Financ. Econ. 117, 470–488.

Fama, E.F., French, K.R., 2018. Choosing factors. J. Financ. Econ. 128, 234–252.

Fama, E.F., MacBeth, J.D., 1973. Risk, return, and equilibrium: empirical tests. J. Polit. Econ. 81, 607–636.

Ferson, W.E., Foerster, S.R., 1994. Finite sample properties of the generalized methods of moments in tests of conditional asset pricing models. J. Financ. Econ. 36, 29–55.

Ferson, W.E., Harvey, C.R., 1999. Conditioning variables and the cross section of stock returns. J. Financ. 54, 1325–1360.

Frazzini, A., Pedersen, L.H., 2014. Betting against beta. J. Financ. Econ. 111, 1–25.

Gibbons, M.R., Ross, S.A., Shanken, J., 1989. A test of the efficiency of a given portfolio. Econometrica 57, 1121–1152.

Giglio, S., Xiu, D., 2019. Inference on Risk Premia in the Presence of Omitted Factors. Yale University. Unpublished working paper

Green, J., Hand, J.R., Zhang, X.F., 2017. The characteristics that provide independent information about average u.s. monthly stock returns. Rev. Financ. Stud. 30, 4389–4436.

Hall, P., Wilson, S.R., 1991. Two guidelines for bootstrap hypothesis testing. Biometrics 47, 757–762.

Harvey, C.R., 2017. Presidential address: the scientific outlook in financial economics. J. Financ. 72, 1399–1440.

Harvey, C.R., Liu, Y., 2020a. False (and missed) discoveries in financial economics. J. Financ. 75, 2503–2553.

Harvey, C.R., Liu, Y., 2020b. Decomposing the Cross-Section of Individual Stock Returns. Duke University. Unpublished working paper

Harvey, C.R., Liu, Y., Zhu, H., 2016. ... and the cross-section of expected returns. Rev. Financ. Stud. 29, 5–68.

Harvey, C.R., Siddique, A., 2000. Conditional skewness in asset pricing tests. J. Financ. 55, 1263–1295.

Hasler, M., Martineau, C., 2019a. The Dynamic CAPM. University of Texas at Dallas. Unpublished working paper

Hasler, M., Martineau, C., 2019b. Does the CAPM Predict Returns?. University of Texas at Dallas. Unpublished working paper

Herskovic, B., Kelly, B.T., Lustig, H.N., Van Nieuwerburgh, S., 2016. The common factor in idiosyncratic volatility: quantitative asset pricing implications. J. Financ. Econ. 119, 249–283.

Hou, K., Xue, C., Zhang, L., 2015. Digesting anomalies: an investment approach. Rev. Financ. Stud. 28, 650–705.

Hou, K., Xue, C., Zhang, L., 2020. Replicating anomalies. Rev. Financ. Stud. 33, 2019–2133.

Jegadeesh, N., Noh, J., Pukthuanthong, K., Roll, R., Wang, J., 2019. Empirical tests of asset pricing models with individual stocks: resovling the errors-in-variables bias in risk premium estimation. J. Financ. Econ. 133, 273–298.

Jensen, M.C., Black, R., Scholes, M.S., 1972. The Capital Asset Pricing Model: Some Empirical Tests. Harvard University. Unpublished working paper

Jensen, T.I., Kelly, B.T., Pedersen, L.H., 2021. Is There a Replication Crisis in Finance?. Copenhagen Business School Unpublished working paper.

Kan, R., Robotti, C., 2009. Model comparison using the Hansen-Jagannathan distance. Rev. Financ. Stud. 22, 3449–3490.

Kelly, B.T., Pruitt, S., Su, Y., 2019. Characteristics are covariances: a unified model of risk and return. J. Financ. Econ. 134, 501–524.

Kosowski, R., Timmermann, A., Wermers, R., White, H., 2006. Can mutual fund "stars" really pick stocks? New evidence from a bootstrap analysis. J. Financ. 61, 2551–2595.

Kozak, S., Nagel, S., Santosh, S., 2018. Interpreting factor models. J. Financ. 73, 1183–1223.

Ledoit, O., Wolf, W., 2003. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. J. Empir. Financ. 10, 603–621.

Lewellen, J., Nagel, S., Shanken, J., 2010. A skeptical appraisal of asset pricing tests. J. Financ. Econ. 96, 175–194.

Lin, X., Palazzo, B., Yang, F., 2020. The risks of old capital age: Asset pricing implications of technology adoption. J. Monet. Econ. 115, 145–161.

Lo, A.W., MacKinlay, A.C., 1990. Data-snooping biases in tests of financial asset pricing models. Rev. Financ. Stud. 3, 431–467.

Ludvigson, S.C., Ng, S., 2009. Macro factors in bond risk premia. Rev. Financ. Stud. 22, 5027–5067.

MacKinlay, A.C., 1987. On multivariate tests of the CAPM. J. Financ. Econ. 18, 341–371.

McLean, R.D., Pontiff, J., 2016. Does academic research destroy stock return predictability? J. Financ. 71, 5–32.

Novy-Marx, R., 2013. The other side of value: the gross profitability premium. J. Financ. Econ. 108, 1–28.

Novy-Marx, R., Velikov, M., 2016. A taxonomy of anomalies and their trading costs. Rev. Financ. Stud. 29, 104–147.

Pástor, L., Stambaugh, R.F., 2003. Liquidity risk and expected stock returns. J. Polit. Econ. 111, 642–685.

Plyakha, Y., Uppal, R., Vilkov, G., 2016. Equal or value weighting? Implications for asset-pricing tests. University of Luxembourg. Unpublished working paper.

Politis, D., Romano, J., 1994. The stationary bootstrap. J. Am. Stat. Assoc. 89, 1303–1313.

Pukthuanthong, K., Roll, R., Subrahmanyam, A., 2019. A protocol for factor identification. Rev. Financ. Stud. 32, 1573–1607.

Romano, J.P., Wolf, M., 2005. Stepwise multiple testing as formalized data snooping. Ecta 73, 1237–1282.

Shanken, J., 1992. On the estimation of beta-pricing model. Rev. Financ. Stud. 5, 1–33.

Sharpe, W.F., 1964. Capital asset prices: a theory of market equilibrium under conditions of risk. J. Financ. 19, 425–442.

Sullivan, S., Timmermann, A., White, H., 1999. Data-snooping, technical trading rule performance, and the bootstrap. J. Financ. 54, 1647–1691.

Treynor, J.L., Black, B., 1973. How to use security analysis to improve portfolio selection. J. Bus. 46, 66–86.

White, H.H., 2000. A reality check for data snooping. Econometrica 68, 1097–1126.