# Lucky Factors
# Online Appendix

**Campbell R. Harvey**

*Duke University, Durham, NC 27708 USA*
*National Bureau of Economic Research, Cambridge, MA 02138 USA*

**Yan Liu**[*]

*Purdue University, West Lafayette, IN 47906 USA*

[*] Current Version: August 12, 2020. First posted on SSRN: November 20, 2014. Previously circulated under the title "How Many Factors?" and "Incremental Factors". Send correspondence to: Yan Liu, Krannert School of Management, Purdue University, West Lafayette, IN 47906. Phone: +1 919.428.1118, E-mail: liu2746@purdue.edu . The full version of the paper is available at http://ssrn.com/abstract=2528780 .

# 1 Summary

This appendix provides additional results to the main paper, “Lucky Factors”, by Campbell R. Harvey and Yan Liu. It is organized as follows.

Appendix IA considers time-series dependency in the data. It first describes the block bootstrap method that we use to adjust for time-series dependency and then presents the testing results, controlling for time-series dependency.

Appendix IB considers several alternative test assets. Appendix IB.1 investigates the impacts of small stocks on our results and presents results that exclude small-cap stocks. Appendix IB.2 presents testing results based on Fama-French 49 industry portfolios. Appendix IB.3 presents testing results based on Fama-French 25 size and book-to-market sorted portfolios and under the value weighted test statistics.

Appendix IC presents testing results that adjust for infrequent trading for some stocks.

Appendix ID describes further details in the bootstrap implementation and shows how to apply our idea to Fama-MacBeth regressions.

Appendix IE relates our approach to five stands of literature: IE.1 for nonparametric bootstrap, IE.2 for factor interpretations, IE.3 for multiple hypothesis testing, IE.4 for sequential tests, and IE.5 for the GRS test, spanning regressions, and a SDF interpretation.

Appendix IF describes and interprets test results based on popular portfolio sorts.

Appendix IG discusses several related issues to our testing framework, including time-varying factor loadings, stepwise model selection, spurious factors, and factor model uncertainty and model misspecification.

Appendix IH (FAQ) answers several questions that the readers may want to ask.

# A Time Dependence

## A.1 Block Bootstrap

Our block bootstrap follows the so-called stationary bootstrap proposed by Politis and Romano (1994) and subsequently applied by White (2000) and Sullivan, Timmermann and White (1999). The stationary bootstrap applies to a strictly stationary and weakly dependent time-series to generate a pseudo time series that is stationary. The stationary bootstrap allows us to resample blocks of the original data, with the length of the block being random and following a geometric distribution with a mean of $1/q$. Therefore, the smoothing parameter $q$ controls the average length of the blocks. A small $q$ (i.e., on average long blocks) is needed for data with strong dependence and a large $q$ (i.e., on average short blocks) is appropriate for data with little dependence. We describe the details of the algorithm in this section.

Suppose the set of time indices for the original data is $1, 2, \ldots, T$. For each bootstrapped sample, our goal is to generate a new set of time indices $\{\theta(t)\}_{t=1}^{T}$. Following Politis and Romano (1994), we first need to choose a smoothing parameter $q$ that can be thought of as the reciprocal of the average block length. The conditions that $q = q_n$ needs to satisfies are:

$$0 < q_n \leq 1, q_n \to 0, nq_n \to \infty.$$

Given this smoothing parameter, we follow the following steps to generate the new set of time indices for each bootstrapped sample:

- Step I. Set $t = 1$ and draw $\theta(1)$ independently and uniformly from $1, 2, \ldots, T$.
- Step II. Move forward one period by setting $t = t+1$. Stop if $t > T$. Otherwise, independently draw a uniformly distributed random variable $U$ on the unit interval.
  1. If $U < q$, draw $\theta(t)$ independently and uniformly from $1, 2, \ldots, T$.
  2. Otherwise (i.e., $U \geq q$), set $\theta(t) = \theta(t-1) + 1$ if $\theta(t) \leq T$ and $\theta(t) = 1$ if $\theta(t) > T$.
- Step III. Repeat step II.

## A.2 Results under Time Dependence

We set $q = 12$ to allow a long enough window for time-series dependency to die off. Table A.21 and A.22 show the results under the equally weighted and the value weighted test statistics, respectively. Compared to Table 3 and 4 in the main paper, allowing for time-series dependency seems to make *smb* less significant. But overall, time-series dependency does not seem to have a substantial impact on our results.

Table A.21: **Individual Stocks as Test Assets, Equally Weighted Scaled Intercepts, Controlling for Time-Series Dependency**

Test results on 14 risk factors using equally weighted individual stocks. (See Table 1 for the definitions of risk factors). We use individual stocks from CRSP that cover the 1968–2012 period to test 14 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 2.2. 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 inference, we rely on block bootstrap with the block size parameter set at $q = 12$, as described in Appendix A.1 of the on-line appendix.

| 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** | [-0.117] | (0.001) | **-0.206** | [-0.119] | (0.002) | | | | | | |
| ***smb*** | -0.081 | [-0.077] | (0.045) | -0.109 | [-0.100] | (0.047) | **-0.041** | [-0.047] | (0.091) | **-0.062** | [-0.064] | (0.077) |
| ***hml*** | 0.088 | [-0.027] | (0.986) | 0.108 | [-0.028] | (0.974) | -0.021 | [-0.023] | (0.107) | -0.047 | [-0.026] | (0.007) |
| ***mom*** | 0.091 | [-0.034] | (1.000) | 0.110 | [-0.044] | (1.000) | 0.070 | [-0.007] | (1.000) | 0.089 | [-0.015] | (1.000) |
| ***skew*** | -0.008 | [-0.028] | (0.332) | -0.002 | [-0.043] | (0.530) | -0.004 | [-0.013] | (0.205) | -0.003 | [-0.014] | (0.362) |
| ***psl*** | 0.011 | [-0.018] | (0.968) | 0.002 | [-0.026] | (0.739) | 0.001 | [-0.007] | (0.357) | -0.003 | [-0.012] | (0.228) |
| ***roe*** | 0.163 | [-0.049] | (0.987) | 0.187 | [-0.063] | (0.995) | 0.142 | [-0.018] | (0.984) | 0.180 | [-0.028] | (0.987) |
| ***ia*** | 0.264 | [-0.034] | (1.000) | 0.291 | [-0.039] | (0.999) | 0.027 | [-0.010] | (0.973) | 0.015 | [-0.012] | (0.972) |
| ***qmj*** | 0.316 | [-0.089] | (0.983) | 0.358 | [-0.107] | (0.988) | 0.149 | [-0.029] | (0.993) | 0.193 | [-0.040] | (0.995) |
| ***bab*** | -0.006 | [-0.038] | (0.562) | -0.049 | [-0.047] | (0.042) | 0.018 | [-0.013] | (0.972) | -0.014 | [-0.020] | (0.087) |
| ***gp*** | 0.017 | [-0.007] | (0.617) | 0.030 | [-0.015] | (0.682) | 0.023 | [-0.001] | (0.968) | 0.017 | [-0.004] | (0.886) |
| ***cma*** | 0.176 | [-0.035] | (1.000) | 0.199 | [-0.037] | (0.991) | -0.012 | [-0.011] | (0.040) | -0.031 | [-0.016] | (0.006) |
| ***rmw*** | 0.116 | [-0.027] | (0.992) | 0.137 | [-0.035] | (0.982) | 0.040 | [-0.015] | (0.997) | 0.048 | [-0.019] | (0.957) |
| ***civ*** | -0.096 | [-0.055] | (0.013) | -0.130 | [-0.070] | (0.009) | -0.018 | [-0.019] | (0.062) | -0.049 | [-0.029] | (0.017) |
| | | multiple test | | | multiple test | | | multiple test | | | multiple test | |
| | *min* | **[-0.126]** | **(0.003)** | | **[-0.140]** | **(0.002)** | *min* | **[-0.050]** | **(0.096)** | | **[-0.065]** | **(0.087)** |
| 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** | [-0.019] | (0.063) | **-0.040** | [-0.021] | (0.006) | | | | | | |
| ***mom*** | 0.055 | [-0.004] | (1.000) | 0.076 | [-0.007] | (1.000) | 0.026 | [-0.006] | (1.000) | 0.046 | [-0.008] | (1.000) |
| ***skew*** | -0.013 | [-0.015] | (0.087) | -0.015 | [-0.017] | (0.063) | **0.006** | [-0.002] | (0.529) | **-0.001** | [-0.003] | (0.676) |
| ***psl*** | 0.011 | [-0.003] | (0.927) | 0.016 | [-0.006] | (0.945) | 0.010 | [-0.003] | (0.986) | 0.007 | [-0.008] | (0.794) |
| ***roe*** | 0.058 | [-0.008] | (0.996) | 0.074 | [-0.009] | (0.988) | 0.072 | [-0.006] | (0.967) | 0.080 | [-0.010] | (0.993) |
| ***ia*** | 0.020 | [-0.015] | (0.980) | 0.008 | [-0.019] | (0.708) | 0.038 | [-0.004] | (0.978) | 0.051 | [-0.006] | (0.975) |
| ***qmj*** | 0.052 | [-0.009] | (0.991) | 0.061 | [-0.010] | (0.970) | 0.128 | [-0.003] | (0.993) | 0.137 | [-0.007] | (0.989) |
| ***bab*** | 0.016 | [-0.014] | (0.893) | -0.014 | [-0.016] | (0.082) | 0.045 | [-0.002] | (0.979) | 0.040 | [-0.004] | (0.990) |
| ***gp*** | 0.022 | [-0.003] | (0.992) | 0.020 | [-0.008] | (0.939) | 0.059 | [-0.002] | (0.988) | 0.055 | [-0.004] | (0.978) |
| ***cma*** | 0.001 | [-0.016] | (0.468) | -0.009 | [-0.018] | (0.176) | 0.022 | [-0.001] | (0.991) | 0.023 | [-0.003] | (0.963) |
| ***rmw*** | -0.009 | [-0.021] | (0.184) | -0.016 | [-0.021] | (0.083) | 0.036 | [-0.003] | (1.000) | 0.043 | [-0.004] | (0.992) |
| ***civ*** | 0.014 | [-0.008] | (0.955) | 0.003 | [-0.012] | (0.606) | 0.015 | [-0.009] | (0.969) | 0.016 | [-0.013] | (0.992) |
| | | multiple test | | | multiple test | | | multiple test | | | multiple test | |
| | *min* | **[-0.026]** | **(0.145)** | | **[-0.030]** | **(0.011)** | *min* | **-0.010** | **(0.982)** | | **-0.019** | **(0.958)** |

Table A.22: **Individual Stocks as Test Assets, Value Weighted Scaled Intercepts, Controlling for Time-Series Dependency**

Test results on 14 risk factors using value weighted individual stocks. (See Table 1 for the definitions of risk factors). We use individual stocks from CRSP that cover the 1968–2012 period to test 14 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 2.2. The metric (i.e., $SI_{vw}$), which measures the difference in value weighted scaled absolute regression intercept, is defined in Section 3.5.2. For inference, we rely on block bootstrap with the block size parameter set at $q = 12$, as described in Appendix A.1. of the on-line appendix.

| 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** | [-0.216] | (0.000) | | | | | | |
| ***smb*** | -0.059 | [-0.045] | (0.019) | 0.018 | [-0.056] | (0.852) | 0.076 | [-0.035] | (0.996) |
| ***hml*** | 0.144 | [-0.067] | (0.983) | -0.038 | [-0.055] | (0.129) | -0.016 | [-0.076] | (0.675) |
| ***mom*** | 0.153 | [-0.075] | (0.993) | 0.130 | [-0.019] | (1.000) | 0.125 | [-0.030] | (1.000) |
| ***skew*** | -0.027 | [-0.041] | (0.068) | -0.044 | [-0.035] | (0.014) | -0.020 | [-0.032] | (0.171) |
| ***psl*** | 0.035 | [-0.029] | (0.976) | 0.016 | [-0.012] | (0.990) | 0.034 | [-0.027] | (0.998) |
| ***roe*** | 0.105 | [-0.048] | (0.991) | -0.079 | [-0.048] | (0.009) | 0.038 | [-0.017] | (0.983) |
| ***ia*** | 0.382 | [-0.065] | (0.980) | -0.042 | [-0.041] | [0.027] | 0.078 | [-0.054] | (0.989) |
| ***qmj*** | 0.363 | [-0.101] | (0.977) | **-0.149** | [-0.063] | (0.005) | | | |
| ***bab*** | -0.048 | [-0.038] | (0.008) | -0.088 | [-0.037] | (0.004) | **-0.026** | [-0.043] | (0.150) |
| ***gp*** | -0.082 | [-0.041] | (0.005) | -0.037 | [-0.045] | (0.078) | -0.022 | [-0.047] | (0.222) |
| ***cma*** | 0.314 | [-0.086] | (0.996) | -0.052 | [-0.033] | (0.017) | 0.019 | [-0.045] | (0.985) |
| ***rmw*** | 0.045 | [-0.010] | (0.903) | -0.146 | [-0.065] | (0.007) | 0.053 | [-0.035] | (0.991) |
| ***civ*** | -0.115 | [-0.048] | (0.003) | 0.035 | [-0.011] | (0.985) | -0.017 | [-0.025] | (0.133) |
| | | multiple test | | | multiple test | | | multiple test | |
| | *min* | **[-0.216]** | **(0.000)** | | **[-0.081]** | **(0.008)** | | **[-0.076]** | **(0.727)** |

# B Alternative Test Assets

## B.1 Dropping Small-Cap Stocks

To mitigate the impact of small-cap stocks, we drop small-cap stocks and rerun our analysis. In particular, at the beginning of each year, we sort the cross-section of individual stocks based on market capitalization and drop stocks in the lowest decile. We redo our tests using the remaining stocks. Table B.11 and B.12 show the results under the equally weighted and the value weighted test statistics, respectively.

Compared to Table 3 in the main paper, under the equally weighted test statistics, focusing on relatively large-cap stocks makes the impact of the market factor larger and the impact of other significant factors smaller. This is not surprising since anomaly returns are usually substantially smaller if we exclude small stocks. Under the value weighted test statistics, our results in Table B.12 are similar to those of Table 4, suggesting that the value weighted test statistics are more robust to the size cutoff that determines our return panel.

Table B.11: **Individual Stocks as Test Assets, Equally Weighted Scaled Intercepts, Excluding Small-Cap Stocks**

Test results on 14 risk factors using equally weighted individual stocks. (See Table 1 for the definitions of risk factors). We use relatively large-cap stocks from CRSP that cover the 1968–2012 period to test 14 risk factors. In particular, at the beginning of each year, we sort the cross-section of stocks based on market capitalization and drop stocks in the lowest decile. We apply our tests to the remaining 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 2.2. The two metrics (i.e., $SI_{ew}^{m}$ and $SI_{ew}^{med}$), which measure the difference in equally weighted scaled mean/median absolute regression intercept, are defined in Section 3.2.

Panel A: Baseline = No factor

| | | single test | | | single test | |
|---|---|---|---|---|---|---|
| Factor | $SI_{ew}^{m}$ | 5th-percentile | $p$-value | $SI_{ew}^{med}$ | 5th-percentile | $p$-value |
| ***mkt*** | **-0.325** | [-0.243] | (0.000) | **-0.345** | [-0.257] | (0.000) |
| ***smb*** | -0.058 | [-0.068] | (0.072) | -0.085 | [-0.080] | (0.028) |
| ***hml*** | 0.163 | [-0.030] | (0.968) | 0.197 | [-0.038] | (0.979) |
| ***mom*** | 0.093 | [-0.056] | (1.000) | 0.096 | [-0.070] | (1.000) |
| ***skew*** | -0.010 | [-0.034] | (0.247) | 0.012 | [-0.047] | (0.716) |
| ***psl*** | 0.040 | [-0.036] | (0.993) | 0.047 | [-0.051] | (0.989) |
| ***roe*** | 0.109 | [-0.073] | (0.974) | 0.115 | [-0.096] | (0.997) |
| ***ia*** | 0.402 | [-0.046] | (1.000) | 0.473 | [-0.049] | (0.992) |
| ***qmj*** | 0.416 | [-0.097] | (0.990) | 0.501 | [-0.122] | (0.974) |
| ***bab*** | -0.014 | [-0.052] | (0.381) | -0.060 | [-0.070] | (0.087) |
| ***gp*** | -0.024 | [-0.043] | (0.237) | -0.010 | [-0.052] | (0.478) |
| ***cma*** | 0.282 | [-0.063] | (1.000) | 0.326 | [-0.074] | (0.986) |
| ***rmw*** | 0.112 | [-0.029] | (0.988) | 0.144 | [-0.038] | (0.990) |
| ***civ*** | -0.112 | [-0.052] | (0.004) | -0.151 | [-0.055] | (0.001) |
| | | multiple test | | | multiple test | |
| | *min* | **[-0.243]** | **(0.000)** | | **[-0.257]** | **(0.000)** |

Panel B: Baseline = ***mkt***

| | | single test | | | single test | |
|---|---|---|---|---|---|---|
| Factor | $SI_{ew}^{m}$ | 5th-percentile | $p$-value | $SI_{ew}^{med}$ | 5th-percentile | $p$-value |
| ***mkt*** | | | | | | |
| ***smb*** | **-0.017** | [-0.032] | (0.294) | **-0.055** | [-0.041] | (0.008) |
| ***hml*** | 0.004 | [-0.036] | (0.800) | -0.025 | [-0.042] | (0.221) |
| ***mom*** | 0.057 | [-0.010] | (1.000) | 0.052 | [-0.020] | (1.000) |
| ***skew*** | 0.001 | [-0.014] | (0.489) | 0.015 | [-0.020] | (0.788) |
| ***psl*** | 0.005 | [-0.014] | (0.702) | -0.013 | [-0.023] | (0.115) |
| ***roe*** | 0.052 | [-0.021] | (0.993) | 0.058 | [-0.024] | (0.971) |
| ***ia*** | 0.013 | [-0.024] | (0.855) | -0.017 | [-0.037] | (0.279) |
| ***qmj*** | 0.030 | [-0.026] | (0.990) | 0.048 | [-0.033] | (0.983) |
| ***bab*** | 0.039 | [-0.015] | (0.994) | -0.011 | [-0.024] | (0.170) |
| ***gp*** | 0.034 | [-0.021] | (0.980) | 0.021 | [-0.030] | (0.952) |
| ***cma*** | -0.009 | [-0.017] | (0.267) | -0.048 | [-0.031] | (0.003) |
| ***rmw*** | 0.001 | [-0.034] | (0.734) | -0.011 | [-0.039] | (0.494) |
| ***civ*** | 0.019 | [-0.0018] | (0.979) | 0.003 | [-0.024] | (0.666) |
| | | multiple test | | | multiple test | |
| | *min* | **[-0.040]** | **(0.661)** | | **[-0.051]** | **(0.017)** |

Panel C: Baseline = ***mkt+smb***

| | | single test | | | single test | |
|---|---|---|---|---|---|---|
| Factor | $SI_{ew}^{m}$ | 5th-percentile | $p$-value | $SI_{ew}^{med}$ | 5th-percentile | $p$-value |
| ***mkt*** | | | | | | |
| ***smb*** | | | | | | |
| ***hml*** | 0.015 | [-0.029] | (0.971) | 0.020 | [-0.037] | (0.918) |
| ***mom*** | 0.054 | [-0.009] | (1.000) | 0.080 | [-0.015] | (1.000) |
| ***skew*** | **-0.009** | [-0.015] | (0.097) | 0.000 | [-0.020] | (0.457) |
| ***psl*** | 0.009 | [-0.012] | (0.895) | 0.023 | [-0.018] | (0.963) |
| ***roe*** | 0.060 | [-0.014] | (0.991) | 0.094 | [-0.025] | (0.976) |
| ***ia*** | 0.026 | [-0.021] | (0.977) | 0.023 | [-0.031] | (0.930) |
| ***qmj*** | 0.071 | [-0.026] | (0.999) | 0.118 | [-0.030] | (0.991) |
| ***bab*** | 0.029 | [-0.018] | (0.943) | 0.018 | [-0.026] | (0.785) |
| ***gp*** | 0.027 | [-0.018] | (1.000) | 0.035 | [-0.027] | (0.983) |
| ***cma*** | 0.001 | [-0.020] | (0.676) | **-0.009** | [-0.028] | (0.234) |
| ***rmw*** | 0.017 | [-0.021] | (0.986) | 0.036 | [-0.038] | (0.971) |
| ***civ*** | 0.019 | [-0.011] | (0.958) | 0.018 | [-0.027] | (0.896) |
| | | multiple test | | | multiple test | |
| | *min* | **[-0.036]** | **(0.831)** | | **[-0.046]** | **(0.897)** |

Table B.12: **Individual Stocks as Test Assets, Value Weighted Scaled Intercepts, Excluding Small-Cap Stocks**

Test results on 14 risk factors using value weighted individual stocks. (See Table 1 for the definitions of risk factors). We use relatively large-cap stocks from CRSP that cover the 1968–2012 period to test 14 risk factors. In particular, at the beginning of each year, we sort the cross-section of stocks based on market capitalization and drop stocks in the lowest decile. We apply our tests to the remaining 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 2.2. The metric (i.e., $SI_{vw}$), which measures the difference in value weighted scaled absolute regression intercept, is defined in Section 3.5.2.

| Panel A: Baseline = No factor | | | | Panel B: Baseline = ***mkt*** | | | Panel C: Baseline = ***mkt+qmj*** | | |
|---|---|---|---|---|---|---|---|---|---|
| | | single test | | | single test | | | single test | |
| Factor | $SI_{vw}$ | 5*th*-percentile | *p*-value | $SI_{vw}$ | 5th-percentile | *p*-value | $SI_{vw}$ | 5th-percentile | *p*-value |
| ***mkt*** | **-0.468** | [-0.267] | (0.000) | | | | | | |
| ***smb*** | -0.045 | [-0.043] | (0.031) | 0.033 | [-0.070] | (0.911) | 0.167 | [-0.040] | (0.987) |
| ***hml*** | 0.162 | [-0.072] | (0.995) | -0.007 | [-0.051] | (0.534) | 0.039 | [-0.074] | (0.983) |
| ***mom*** | 0.151 | [-0.067] | (1.000) | 0.125 | [-0.014] | (1.000) | 0.124 | [-0.038] | (1.000) |
| ***skew*** | -0.023 | [-0.043] | (0.169) | -0.040 | [-0.041] | (0.060) | -0.011 | [-0.025] | (0.190) |
| ***psl*** | 0.038 | [-0.026] | (0.975) | 0.022 | [-0.014] | (0.986) | 0.045 | [-0.019] | (0.999) |
| ***roe*** | 0.086 | [-0.051] | (0.997) | -0.124 | [-0.057] | (0.002) | 0.035 | [-0.023] | (0.981) |
| ***ia*** | 0.408 | [-0.085] | (0.983) | -0.019 | [-0.054] | (0.029) | 0.144 | [-0.057] | (0.989) |
| ***qmj*** | 0.371 | [-0.128] | (0.998) | **-0.200** | [-0.086] | (0.001) | | | |
| ***bab*** | -0.037 | [-0.041] | (0.073) | -0.062 | [-0.035] | (0.007) | 0.027 | [-0.036] | (0.975) |
| ***gp*** | -0.091 | [-0.054] | (0.003) | -0.035 | [-0.045] | (0.123) | **-0.040** | [-0.061] | (0.218) |
| ***cma*** | 0.339 | [-0.093] | (0.989) | -0.028 | [-0.043] | (0.120) | 0.070 | [-0.053] | (0.998) |
| ***rmw*** | 0.042 | [-0.016] | (0.945) | -0.160 | [-0.080] | (0.004) | 0.042 | [-0.027] | (0.990) |
| ***civ*** | -0.110 | [-0.055] | (0.006) | 0.061 | [-0.029] | (0.997) | 0.016 | [-0.016] | (0.976) |
| | | multiple test | | | multiple test | | | multiple test | |
| | *min* | **[-0.267]** | **(0.000)** | | **[-0.099]** | **(0.003)** | | **[-0.078]** | **(0.513)** |

## B.2 Fama-French 49 Industry Portfolios

We use Fama-French 49 industry portfolios as test assets. The sample period is from 1972 to 2012. To construct the value weighted test statistics, we use the aggregate market capitalization for each industry portfolio as the value weight.

By using the 49 industry portfolios, the impact of the market factor becomes larger compared to its impact when using individual stocks. This is because the level of noise is lower for sorted portfolios than for individual stocks. However, we are also losing power in identifying true risk factors by using industry portfolios. In particular, neither *smb* nor *hml* is significant under the equally weighted test statistics. The results under value weighting are similar to the results in the main paper.

## Table B.21: **Fama-French 49 Portfolios as Test Assets, Equally Weighted Scaled Intercepts**

Test results on 14 risk factors using Fama-French 49 portfolios, 1972-2012. (See Table 1 for the definitions of risk factors.). The baseline model refers to the model that includes the pre-selected risk factors. We focus on the panel regression model described in Section 2.2. The two metrics (i.e., $SI_{ew}^{m}$ and $SI_{ew}^{med}$), which measure the difference in equally weighted scaled mean/median absolute regression intercepts, are defined in Section 3.2.

| Panel A: Baseline = No factor | | | | | | | Panel B: Baseline = *mkt* | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | single test | | | single test | | | single test | | | single test | |
| Factor | $SI_{ew}^{m}$ | 5$th$-percentile | $p$-value | $SI_{ew}^{med}$ | 5th-percentile | $p$-value | $SI_{ew}^{m}$ | 5th-percentile | $p$-value | $SI_{ew}^{med}$ | 5th-percentile | $p$-value |
| ***mkt*** | **-0.670** | [-0.366] | (0.000) | **-0.757** | [-0.387] | (0.000) | | | | | | |
| ***smb*** | -0.162 | [-0.185] | (0.071) | -0.182 | [-0.218] | (0.084) | 0.061 | [-0.108] | (0.910) | 0.166 | [-0.170] | (0.959) |
| ***hml*** | 0.249 | [-0.122] | (0.994) | 0.273 | [-0.155] | (0.987) | 0.055 | [-0.102] | (0.918) | 0.073 | [-0.174] | (0.846) |
| ***mom*** | 0.286 | [-0.134] | (0.999) | 0.336 | [-0.166] | (0.996) | 0.158 | [-0.037] | (1.000) | 0.476 | [-0.098] | (1.000) |
| ***skew*** | -0.042 | [-0.051] | (0.069) | -0.052 | [-0.076] | (0.090) | -0.126 | [-0.072] | (0.007) | **-0.189** | [-0.137] | (0.015) |
| ***psl*** | 0.028 | [-0.035] | (0.912) | 0.044 | [-0.051] | (0.907) | -0.019 | [-0.026] | (0.097) | -0.002 | [-0.083] | (0.498) |
| ***roe*** | 0.337 | [-0.111] | (1.000) | 0.343 | [-0.144] | (0.998) | -0.084 | [-0.077] | (0.040) | 0.107 | [-0.134] | (0.928) |
| ***ia*** | 0.673 | [-0.164] | (1.000) | 0.682 | [-0.201] | (1.000) | 0.016 | [-0.097] | (0.991) | 0.320 | [-0.152] | (0.999) |
| ***qmj*** | 0.681 | [-0.257] | (0.994) | 0.647 | [-0.282] | (0.975) | -0.089 | [-0.099] | (0.059) | -0.023 | [-0.134] | (0.353) |
| ***bab*** | 0.012 | [-0.039] | (0.774) | -0.065 | [-0.067] | (0.052) | 0.003 | [-0.089] | (0.554) | 0.091 | [-0.149] | (0.880) |
| ***gp*** | -0.131 | [-0.080] | (0.004) | -0.183 | [-0.110] | (0.005) | 0.086 | [-0.058] | (0.999) | 0.134 | [-0.142] | (0.945) |
| ***cma*** | 0.574 | [-0.178] | (1.000) | 0.588 | [-0.213] | (0.997) | 0.019 | [-0.076] | (0.779) | 0.216 | [-0.137] | (0.992) |
| ***rmw*** | 0.143 | [-0.069] | (0.990) | 0.133 | [-0.102] | (0.954) | **-0.145** | [-0.114] | (0.021) | -0.040 | [-0.168] | (0.321) |
| ***civ*** | -0.274 | [-0.144] | (0.001) | -0.339 | [-0.170] | (0.002) | 0.079 | [-0.036] | (0.996) | 0.120 | [-0.109] | (0.962) |
| | | multiple test | | | multiple test | | | multiple test | | | multiple test | |
| | *min* | **[-0.370]** | **(0.000)** | | **[-0.407]** | **(0.000)** | *min* | **[-0.164]** | **(0.075)** | | **[-0.264]** | **(0.179)** |

## Table B.22: **Fama-French 49 Portfolios as Test Assets, Value Weighted Scaled Intercepts**

Test results on 14 risk factors using Fama-French 49 portfolios, 1972-2012. (See Table 1 for the definitions of risk factors.). The baseline model refers to the model that includes the pre-selected risk factors. We focus on the panel regression model described in Section 2.2. The metric (i.e., $SI_{vw}$), which measures the difference in value weighted scaled absolute regression intercept, is defined in Section 3.5.2.

| 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*** | **-0.580** | [-0.307] | (0.000) | | | | | | |
| ***smb*** | -0.087 | [-0.099] | (0.071) | 0.054 | [-0.096] | (0.828) | 0.270 | [-0.060] | (1.000) |
| ***hml*** | 0.183 | [-0.091] | (0.991) | -0.117 | [-0.115] | (0.049) | 0.207 | [-0.140] | (0.997) |
| ***mom*** | 0.203 | [-0.093] | (0.997) | 0.094 | [-0.025] | (0.997) | 0.151 | [-0.059] | (0.998) |
| ***skew*** | -0.034 | [-0.041] | (0.071) | -0.132 | [-0.077] | (0.008) | 0.026 | [-0.062] | (0.913) |
| ***psl*** | 0.023 | [-0.026] | (0.926) | -0.005 | [-0.022] | (0.235) | -0.033 | [-0.029] | (0.041) |
| ***roe*** | 0.184 | [-0.074] | (1.000) | -0.214 | [-0.094] | (0.001) | 0.062 | [-0.040] | (0.995) |
| ***ia*** | 0.468 | [-0.134] | (1.000) | -0.163 | [-0.104] | [0.010] | 0.278 | [-0.107] | (1.000) |
| ***qmj*** | 0.442 | [-0.186] | (0.998) | **-0.295** | [-0.118] | (0.000) | | | |
| ***bab*** | -0.035 | [-0.028] | (0.034) | -0.232 | [-0.112] | (0.000) | 0.165 | [-0.089] | (0.997) |
| ***gp*** | -0.110 | [-0.060] | (0.007) | -0.014 | [-0.055] | (0.880) | **-0.039** | [-0.113] | (0.253) |
| ***cma*** | 0.403 | [-0.134] | (0.998) | -0.199 | [-0.096] | (0.000) | 0.165 | [-0.105] | (0.997) |
| ***rmw*** | 0.064 | [-0.037] | (0.980) | -0.277 | [-0.138] | (0.002) | 0.105 | [-0.087] | (0.992) |
| ***civ*** | -0.175 | [-0.085] | (0.002) | 0.078 | [-0.034] | (0.996) | 0.159 | [-0.039] | (1.000) |
| | | multiple test | | | multiple test | | | multiple test | |
| | *min* | **[-0.308]** | **(0.000)** | | **[-0.173]** | **(0.002)** | | **[-0.165]** | **(0.698)** |

## B.3 Fama-French 25 Portfolios, Value Weighting

We use Fama-French 25 portfolios as test assets. We only examine the value weighted test statistics. Table 2 in the main paper examines the equally weighted test statistics.

Using Fama-French 25 portfolios and under value weighting, *bab* is the next factor identified after the market factor. This contradicts the results under value weighting either for Fama-French 49 portfolios or individual stocks (i.e., Table 4), for which *qmj* is identified as the next factor after the market factor. This contradiction points to the difficulty in interpreting test results based on different sets of portfolios. A certain set of portfolios may be biased towards identifying a certain set of factors. To avoid this bias, we advocate the use of individual stocks.

Table B.31: **Fama-French 25 Portfolios as Test Assets, Value Weighted Scaled Intercepts**

Test results on 14 risk factors using Fama-French 25 portfolio. (See Table 1 for the definitions of risk factors.). The baseline model refers to the model that includes the pre-selected risk factors. We focus on the panel regression model described in Section 2.2. The metric (i.e., $SI_{vw}$), which measures the difference in value weighted scaled absolute regression intercept, is defined in Section 3.5.2.

| 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*** | **-0.754** | [-0.407] | (0.000) | | | | | | |
| ***smb*** | -0.107 | [-0.126] | (0.076) | 0.057 | [-0.193] | (0.881) | 0.024 | [-0.272] | (0.805) |
| ***hml*** | 0.191 | [-0.123] | (0.980) | -0.076 | [-0.370] | (0.404) | 0.779 | [-0.389] | (0.999) |
| ***mom*** | 0.241 | [-0.127] | (0.986) | 0.355 | [-0.102] | (0.998) | **-0.118** | [-0.170] | (0.104) |
| ***skew*** | -0.023 | [-0.053] | (0.157) | -0.258 | [-0.148] | (0.006) | 0.034 | [-0.135] | (0.801) |
| ***psl*** | 0.047 | [-0.050] | (0.922) | -0.033 | [-0.042] | (0.070) | -0.054 | [-0.046] | (0.034) |
| ***roe*** | 0.330 | [-0.130] | (0.997) | 0.349 | [-0.123] | (1.000) | 0.654 | [-0.176] | (1.000) |
| ***ia*** | 0.610 | [-0.168] | (0.998) | 0.274 | [-0.283] | (0.913) | 2.105 | [-0.280] | (0.978) |
| ***qmj*** | 0.724 | [-0.249] | (0.968) | 0.637 | [-0.180] | (0.997) | 1.212 | [-0.251] | (0.992) |
| ***bab*** | 0.028 | [-0.039] | (0.891) | **-0.477** | [-0.204] | (0.000) | | | |
| ***gp*** | -0.044 | [-0.042] | (0.045) | 0.700 | [-0.249] | (0.999) | 0.269 | [-0.272] | (0.958) |
| ***cma*** | 0.515 | [-0.187] | (0.994) | -0.039 | [-0.306] | (0.435) | 1.555 | [-0.291] | (1.000) |
| ***rmw*** | 0.160 | [-0.094] | (0.967) | -0.066 | [-0.134] | (0.128) | 0.123 | [-0.194] | (0.958) |
| ***civ*** | -0.241 | [-0.124] | (0.009) | -0.196 | [-0.086] | (0.002) | 0.097 | [-0.088] | (0.965) |
| | | multiple test | | | multiple test | | | multiple test | |
| | *min* | **[-0.419]** | **(0.000)** | | **[-0.406]** | **(0.017)** | | **[-0.432]** | **(0.686)** |

## B.4 Combining Factors

To examine whether combinations of several factors provide better risk factors, we add two additional factors to the original 14 risk factors. They are $hml_a = (hml + ia + cma)/3$ and $qmj_a = (qmj + roe + rmw)/3$, motivated by the factor correlation matrix shown in the main paper. Table B.41 and B.42 show the results under the equally weighted and the value weighted test statistics, respectively.

Table B.41: **Combining Factors, Equally Weighted Scaled Intercepts**

Test results on 16 risk factors using equally weighted individual stocks. (See Table I for the definitions of 14 risk factors). Besides the factors in Table I, we include two additional factors that are based on combinations of the original 14 factors. They are $hml_a = (hml + ia + cma)/3$ and $qmj_a = (qmj + roe + rmw)/3$. We use individual stocks from CRSP that cover the 1968–2012 period to test 16 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 2.2. 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.

| | Panel A: Baseline = No factor | | | | | | Panel B: Baseline = ***mkt*** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | single test | | | single test | | | single test | | | single test | |
| Factor | $SI^m_{ew}$ | 5$th$-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) |
| $hml_a$ | 0.178 | [-0.029] | (1.000) | 0.196 | [-0.025] | (1.000) | 0.002 | [-0.021] | (0.719) | -0.018 | [-0.028] | (0.135) |
| $qmj_a$ | 0.252 | [-0.025] | (1.000) | 0.294 | [-0.034] | (1.000) | 0.137 | [-0.025] | (1.000) | 0.181 | [-0.034] | (1.000) |
| | | multiple test | | | multiple test | | | multiple test | | | multiple test | |
| | *min* | **[-0.098]** | **(0.003)** | | **[-0.106]** | **(0.001)** | *min* | **[-0.050]** | **(0.093)** | | **[-0.063]** | **(0.041)** |

| | Panel C: Baseline = ***mkt+smb*** | | | | | | Panel D: Baseline = ***mkt*** + ***smb+hml*** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | single test | | | single test | | | single test | | | single test | |
| Factor | $SI^m_{ew}$ | 5$th$-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) |
| $hml_a$ | 0.008 | [-0.016] | (0.760) | -0.013 | [-0.016] | (0.093) | 0.032 | [-0.001] | (0.978) | 0.039 | [-0.005] | (1.000) |
| $qmj_a$ | 0.041 | [-0.010] | (0.993) | 0.047 | [-0.010] | (1.000) | 0.114 | [-0.003] | (1.000) | 0.110 | [-0.007] | (1.000) |
| | | multiple test | | | multiple test | | | multiple test | | | multiple test | |
| | *min* | **[-0.019]** | **(0.091)** | | **[-0.023]** | **(0.005)** | *min* | **[-0.009]** | **(0.981)** | | **[-0.014]** | **(0.974)** |

## Table B.42: **Combining Factors, Value Weighted Scaled Intercepts**

Test results on 16 risk factors using value weighted individual stocks. (See Table I for the definitions of 14 risk factors). Besides the factors in Table I, we include two additional factors that are based on combinations of the original 14 factors. They are $hml_a = (hml + ia + cma)/3$ and $qmj_a = (qmj + roe + rmw)/3$. We use individual stocks from CRSP that cover the 1968–2012 period to test 16 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 2.2. The two metrics (i.e., $SI_{ew}^{m}$ and $SI_{ew}^{med}$), which measure the difference in equally weighted scaled mean/median absolute regression intercept, are defined in Section 3.2.

| 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** | [-0.237] | (0.000) | | | | | | |
| ***smb*** | -0.059 | [-0.055] | (0.043) | 0.018 | [-0.046] | (0.827) | 0.081 | [-0.027] | (0.979) |
| ***hml*** | 0.144 | [-0.059] | (0.976) | -0.038 | [-0.051] | (0.129) | -0.015 | [-0.067] | (0.617) |
| ***mom*** | 0.153 | [-0.063] | (0.997) | 0.130 | [-0.012] | (0.999) | 0.111 | [-0.026] | (1.000) |
| ***skew*** | -0.027 | [-0.054] | (0.160) | -0.044 | [-0.034] | (0.031) | -0.010 | [-0.027] | (0.197) |
| ***psl*** | 0.035 | [-0.021] | (0.971) | 0.016 | [-0.011] | (0.991) | 0.006 | [-0.016] | (0.791) |
| ***roe*** | 0.105 | [-0.047] | (0.992) | -0.079 | [-0.046] | (0.020) | 0.045 | [-0.028] | (0.978) |
| ***ia*** | 0.382 | [-0.086] | (0.981) | -0.042 | [-0.049] | (0.079) | 0.093 | [-0.053] | (0.956) |
| ***qmj*** | 0.363 | [-0.111] | (0.887) | -0.149 | [-0.082] | (0.003) | 0.001 | [-0.018] | (0.873) |
| ***bab*** | -0.048 | [-0.033] | (0.031) | -0.088 | [-0.047] | (0.007) | **-0.021** | [-0.036] | (0.241) |
| ***gp*** | -0.082 | [-0.039] | (0.010) | -0.037 | [-0.040] | (0.078) | -0.012 | [-0.042] | (0.515) |
| ***cma*** | 0.314 | [-0.110] | (0.991) | -0.052 | [-0.035] | (0.029) | 0.057 | [-0.044] | (1.000) |
| ***rmw*** | 0.045 | [-0.012] | (0.931) | -0.146 | [-0.067] | (0.021) | 0.130 | [-0.025] | (0.978) |
| ***civ*** | -0.115 | [-0.062] | (0.003) | 0.035 | [-0.020] | (0.977) | 0.005 | [-0.016] | (0.876) |
| ***$hml_a$*** | 0.294 | [-0.070] | (0.991) | -0.012 | [-0.051] | (0.487) | 0.096 | [-0.060] | (0.987) |
| ***$qmj_a$*** | 0.222 | [-0.072] | (1.000) | **-0.154** | [-0.050] | (0.002) | | | |
| | | multiple test | | | multiple test | | | multiple test | |
| | *min* | **[-0.237]** | **(0.000)** | | **[-0.079]** | **(0.006)** | | **[-0.068]** | **(0.782)** |

# C Liquidity Adjustment

Stocks may be infrequently traded. As a result, a stock's contemporaneous exposure to a risk factor may be insufficient to capture its overall exposure to the risk factor. We therefore include both contemporaneous factors and lagged factors to adjust for infrequent trading. In our testing framework, we treat a contemporaneous factor and its lag as a package. So when we construct the null hypothesis to test the incremental contribution of a new factor to existing factors, we adjust the new factor to not only existing contemporaneous factors, but also their lags.

Table C.1 and C.2 show the results under equal weighting and value weighting, respectively. Compared to Table 3 and 4 in the main paper, under equal weighting, the evidence for *hml* is weaker after controlling for lagged factors. Under value weighting, the results are similar to Table 4.

## Table C.1: **Individual Stocks as Test Assets, Equally Weighted Scaled Intercepts, Controlling for Lagged Factors**

Test results on 14 risk factors using equally weighted individual stocks. (See Table 1 for the definitions of risk factors). We use individual stocks from CRSP that cover the 1968–2012 period to test 14 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 2.2. 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.

| 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*** | **-0.190** | [-0.124] | (0.000) | **-0.215** | [-0.134] | (0.001) | | | | | | |
| ***smb*** | -0.082 | [-0.069] | (0.039) | -0.114 | [-0.92] | (0.028) | **-0.033** | [-0.027] | (0.041) | **-0.052** | [-0.043] | (0.013) |
| ***hml*** | 0.107 | [-0.030] | (0.991) | 0.122 | [-0.039] | (0.980) | -0.007 | [-0.022] | (0.318) | -0.036 | [-0.028] | (0.021) |
| ***mom*** | 0.127 | [-0.032] | (1.000) | 0.146 | [-0.049] | (1.000) | 0.086 | [-0.007] | (1.000) | 0.106 | [-0.014] | (1.000) |
| ***skew*** | -0.003 | [-0.028] | (0.341) | -0.001 | [-0.040] | (0.452) | 0.005 | [-0.008] | (0.385) | 0.004 | [-0.009] | (0.436) |
| ***psl*** | 0.024 | [-0.019] | (0.962) | 0.019 | [-0.029] | (0.860) | 0.012 | [-0.003] | (0.694) | 0.012 | [-0.011] | (0.720) |
| ***roe*** | 0.190 | [-0.072] | (0.979) | 0.222 | [-0.094] | (0.969) | 0.146 | [-0.019] | (0.994) | 0.187 | [-0.027] | (0.988) |
| ***ia*** | 0.361 | [-0.030] | (1.000) | 0.394 | [-0.038] | (0.997) | 0.070 | [-0.004] | (0.988) | 0.053 | [-0.009] | (0.990) |
| ***qmj*** | 0.340 | [-0.090] | (0.991) | 0.395 | [-0.114] | (0.973) | 0.142 | [-0.023] | (0.979) | 0.179 | [-0.034] | (1.000) |
| ***bab*** | 0.039 | [-0.036] | (0.980) | 0.006 | [-0.056] | (0.664) | 0.060 | [-0.003] | (0.972) | 0.033 | [-0.006] | (0.972) |
| ***gp*** | 0.022 | [-0.013] | (0.631) | 0.043 | [-0.009] | (0.871) | 0.020 | [-0.004] | (0.883) | 0.021 | [-0.008] | (0.870) |
| ***cma*** | 0.225 | [-0.042] | (0.996) | 0.245 | [-0.045] | (0.985) | 0.012 | [-0.008] | (0.827) | -0.018 | [-0.013] | (0.017) |
| ***rmw*** | 0.126 | [-0.029] | (0.983) | 0.147 | [-0.040] | (0.989) | 0.030 | [-0.019] | (0.995) | 0.027 | [-0.028] | (0.971) |
| ***civ*** | -0.086 | [-0.048] | (0.003) | -0.122 | [-0.063] | (0.002) | 0.003 | [-0.016] | (0.717) | -0.020 | [-0.020] | (0.054) |
| | | multiple test | | | multiple test | | | multiple test | | | multiple test | |
| | *min* | **[-0.136]** | **(0.001)** | | **[-0.154]** | **(0.003)** | *min* | **[-0.032]** | **(0.056)** | | **[-0.044]** | **(0.016)** |

| Panel C: Baseline = *mkt+smb* | | | | | | |
|---|---|---|---|---|---|---|
| | | single test | | | single test | |
| Factor | $SI^{m}_{ew}$ | $5th$-percentile | $p$-value | $SI^{med}_{ew}$ | 5th-percentile | $p$-value |
| ***mkt*** | | | | | | |
| ***smb*** | | | | | | |
| ***hml*** | 0.009 | [-0.012] | (0.444) | **-0.017** | [-0.019] | (0.052) |
| ***mom*** | 0.078 | [0.000] | (1.000) | 0.097 | [-0.009] | (1.000) |
| ***skew*** | **-0.004** | [-0.009] | (0.246) | 0.001 | [-0.010] | (0.246) |
| ***psl*** | 0.024 | [0.003] | (0.937) | 0.024 | [-0.001] | (0.911) |
| ***roe*** | 0.079 | [-0.004] | (0.989) | 0.078 | [-0.008] | (0.987) |
| ***ia*** | 0.077 | [-0.003] | (0.971) | 0.058 | [-0.006] | (0.993) |
| ***qmj*** | 0.072 | [0.000] | (1.000) | 0.075 | [-0.005] | (0.995) |
| ***bab*** | 0.063 | [-0.001] | (0.853) | 0.034 | [-0.006] | (0.910) |
| ***gp*** | 0.018 | [-0.006] | (0.978) | 0.012 | [-0.009] | (0.558) |
| ***cma*** | 0.031 | [-0.006] | (0.997) | 0.008 | [-0.008] | (0.406) |
| ***rmw*** | 0.010 | [-0.013] | (0.694) | -0.000 | [-0.014] | (0.327) |
| ***civ*** | 0.029 | [-0.005] | (0.980) | 0.023 | [-0.008] | (0.988) |
| | | multiple test | | | multiple test | |
| | *min* | **[-0.017]** | **(0.781)** | | **[-0.021]** | **(0.109)** |

Table C.2: **Individual Stocks as Test Assets, Value Weighted Scaled Intercepts, Controlling for Lagged Factors**

Test results on 14 risk factors using value weighted individual stocks. (See Table 1 for the definitions of risk factors). We use individual stocks from CRSP that cover the 1968–2012 period to test 14 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 2.2. The metric (i.e., $SI_{vw}$), which measures the difference in value weighted scaled absolute regression intercept, is defined in Section 3.5.2.

| 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*** | **-0.434** | [-0.318] | (0.000) | | | | | | |
| ***smb*** | -0.068 | [-0.073] | (0.061) | 0.002 | [-0.062] | (0.663) | 0.040 | [-0.048] | (1.000) |
| ***hml*** | 0.148 | [-0.066] | (0.987) | -0.030 | [-0.049] | (0.219) | 0.021 | [-0.053] | (0.991) |
| ***mom*** | 0.165 | [-0.058] | (1.000) | 0.142 | [-0.020] | (0.999) | 0.150 | [-0.025] | (0.968) |
| ***skew*** | -0.021 | [-0.040] | (0.211) | -0.032 | [-0.044] | (0.117) | **-0.005** | [-0.018] | (0.342) |
| ***psl*** | 0.035 | [-0.026] | (0.939) | 0.029 | [-0.016] | (0.993) | 0.031 | [-0.022] | (0.985) |
| ***roe*** | 0.112 | [-0.066] | (0.991) | -0.085 | [-0.054] | (0.002) | 0.082 | [-0.020] | (1.000) |
| ***ia*** | 0.453 | [-0.076] | (0.991) | -0.012 | [-0.051] | (0.510) | 0.082 | [-0.043] | (0.990) |
| ***qmj*** | 0.332 | [-0.133] | (0.983) | **-0.134** | [-0.063] | (0.003) | | | |
| ***bab*** | -0.035 | [-0.045] | (0.107) | -0.052 | [-0.045] | (0.041) | 0.050 | [-0.034] | (0.983) |
| ***gp*** | -0.114 | [-0.052] | (0.006) | -0.057 | [-0.041] | (0.017) | 0.027 | [-0.040] | (0.981) |
| ***cma*** | 0.345 | [-0.100] | (0.966) | -0.028 | [-0.042] | (0.142) | 0.007 | [-0.053] | (0.963) |
| ***rmw*** | 0.055 | [-0.024] | (0.983) | -0.137 | [-0.068] | (0.008) | 0.014 | [-0.033] | (0.941) |
| ***civ*** | -0.100 | [-0.067] | (0.018) | 0.043 | [-0.024] | (1.000) | 0.078 | [-0.017] | (0.993) |
| | | multiple test | | | multiple test | | | multiple test | |
| | *min* | **[-0.318]** | **(0.002)** | | **[-0.078]** | **(0.006)** | | **[-0.062]** | **(0.971)** |

# D Alternative Testing Frameworks

We first illustrate our bootstrapping-based framework within the context of predictive regressions. We then adapt it to the Fama-Macbeth cross-sectional regressions.

## D.1 Predictive Regressions

Suppose we have a $T \times 1$ vector $Y$ of returns that we want to predict and a $T \times M$ matrix $X$ that includes the time-series of $M$ right-hand side variables, i.e., column $i$ of matrix $X$ $(X_i)$ gives the time-series of variable $i$. Our goal is to select a subset of the $M$ regressors to form the "best" predictive regression model. Suppose we measure the goodness-of-fit of a regression model by the summary statistic $\Psi$. Our framework permits the use of an arbitrary performance measure $\Psi$, e.g., $R^2$, $t$-statistic or F-statistic. This feature stems from our use of the bootstrap method, which does not rely on asymptotic approximations of the summary statistics to construct the test. In contrast, FSW need the finite-sample distribution of the $R^2$ statistic to construct their test. To ease the presentation, we describe our approach with the usual regression $R^2$.

Our bootstrap-based selection procedure consists of three major steps:

***Step I. Orthogonalization Under the Null***

Suppose we already selected $k$ $(0 \leq k < M)$ variables and want to test if there exists another significant predictor, that is, we are testing among the rest $M - k$ candidate variables, i.e., $\{X_{k+j}, j = 1, \ldots, M - k\}$. Our null hypothesis is that none of these candidate variables provides additional explanatory power of $Y$, following White (2000) and FSW. The goal of this step is to modify the data matrix $X$ such that this null hypothesis appears to be true in-sample.

To achieve this, we first project $Y$ onto the group of pre-selected variables and obtain the projection residual vector $Y^{e,k}$. This residual vector contains information that cannot be explained by pre-selected variables. We then orthogonalize the $M - k$ candidate variables so that no candidate variable is correlated with $Y^{e,k}$, for the entire sample. In particular, we individually project $X_{k+1}, X_{k+2}, \ldots, X_M$ onto $Y^{e,k}$ and obtain the projection residuals $X^e_{k+1}, X^e_{k+2}, \ldots, X^e_M$, i.e.,

$$X_{k+j} = c_j + d_j Y^{e,k} + X^e_{k+j}, \quad j = 1, \ldots, M - k, \tag{D.1}$$

where $c_j$ is the intercept, $d_j$ is the slope and $X^e_{k+j}$ is the residual vector. By construction, these residuals have an in-sample correlation of zero with $Y^{e,k}$. Therefore, they appear to be independent of $Y^{e,k}$ if joint normality is assumed between $X$ and $Y^{e,k}$.

This is similar to the simulation approach in FSW, in which artificially generated independent regressors are used to quantify the effect of the multiple testing. Our approach is different from FSW because we use real data. In addition, we use bootstrap or block bootstrap to approximate the empirical distribution of test statistics.

We achieve the same goal as FSW while losing as little information as possible for the dependence structure among the regressors. In particular, our orthogonalization guarantees that the $M - k$ orthogonalized candidate variables are uncorrelated with $Y^{e,k}$ in-sample.[1] This resembles the independence requirement between the simulated regressors and the left-hand side variables in FSW. Our approach is distribution free and maintains as much information as possible among the regressors. We simply purge $Y^{e,k}$ out of each of the candidate variables and therefore keep all the distributional information among the variables that is not linearly related to $Y^{e,k}$ intact. For instance, the tail dependency among all the variables — both pre-selected and candidate — is preserved. This is important because higher moment dependence may have a dramatic impact on the test statistics in finite samples.[2]

A similar idea has been applied to the recent literature on mutual fund performance. In particular, Kosowski et al. (2006) and Fama and French (2010) subtract the in-sample fitted alphas from fund returns, thereby creating "pseudo" funds that exactly generate a mean return of zero in-sample. Analogously, we orthogonalize candidate regressors such that they exactly have a correlation of zero with what is left to explain in the left-hand side variable, i.e., $Y^{e,k}$.

***Step II. Bootstrap***

Let us arrange the pre-selected variables into $X^s = [X_1, X_2, \ldots, X_k]$ and the orthogonalized candidate variables into $X^e = [X^e_{k+1}, X^e_{k+2}, \ldots, X^e_M]$. Notice that for both the residual response vector $Y^{e,k}$ and the two regressor matrices $X^s$ and $X^e$, rows denote time periods and columns denote variables. We bootstrap the time periods (i.e., rows) to generate the empirical distributions of the summary statistics for different regression models. In particular, for each draw of the time index $t^b = [t^b_1, t^b_2, \ldots, t^b_T]'$, let the corresponding left-hand side and right variables be $Y^{eb}$, $X^{sb}$, and $X^{eb}$.

The diagram below illustrates how we bootstrap. Suppose we have five periods, one pre-selected variable $X^s$, and one candidate variable $X^e$. The original time index is given by $[t_1 = 1, t_2 = 2, t_3 = 3, t_4 = 4, t_5 = 5]'$. By sampling with replacement,

[1] In fact, the zero correlation between the candidate variables and $Y^{e,k}$ not only holds in-sample, but also in the bootstrapped population provided that each sample period has an equal chance of being sampled in the bootstrapping, which is true in an independent bootstrap. When we use a stationary bootstrap to take time dependency into account, this is no longer true as samples on the boundary time periods are sampled less frequently. But we should expect this correlation to be small for a long enough sample as the boundary periods are a small fraction of the total time periods.

[2] See Adler, Feldman and Taqqu (1998) for how distributions with heavy tails affect standard statistical inference.

one possible realization of the time index for the bootstrapped sample is $t^b = [t^b_1 = 3, t^b_2 = 2, t^b_3 = 4, t^b_4 = 3, t^b_5 = 1]'$. The diagram shows how we transform the original data matrix into the bootstrapped data matrix based on the new time index.[3]

$$[Y^{e,k}, X^s, X^e] = \underbrace{\begin{bmatrix} y^e_1 & x^s_1 & x^e_1 \\ y^e_2 & x^s_2 & x^e_2 \\ y^e_3 & x^s_3 & x^e_3 \\ y^e_4 & x^s_4 & x^e_4 \\ y^e_5 & x^s_5 & x^e_5 \end{bmatrix}}_{\text{Original data matrix}} \begin{pmatrix} t_1 = 1 \\ t_2 = 2 \\ t_3 = 3 \\ t_4 = 4 \\ t_5 = 5 \end{pmatrix} \Rightarrow \begin{pmatrix} t^b_1 = 3 \\ t^b_2 = 2 \\ t^b_3 = 4 \\ t^b_4 = 3 \\ t^b_5 = 1 \end{pmatrix} \underbrace{\begin{bmatrix} y^e_3 & x^s_3 & x^e_3 \\ y^e_2 & x^s_2 & x^e_2 \\ y^e_4 & x^s_4 & x^e_4 \\ y^e_3 & x^s_3 & x^e_3 \\ y^e_1 & x^s_1 & x^e_1 \end{bmatrix}}_{\text{Bootstrapped data matrix}} = [Y^{eb}, X^{sb}, X^{eb}]$$

Returning to the general case with $k$ pre-selected variables and $M - k$ candidate variables, we bootstrap and then run $M - k$ regressions. Each of these regressions involves the projection of $Y^{eb}$ onto a candidate variable from the data matrix $X^{eb}$. Let the associated summary statistics be $\Psi^{k+1,b}$, $\Psi^{k+2,b}$, $\ldots$, $\Psi^{M,b}$, and let the maximum among these summary statistics be $\Psi^b_I$, i.e.,

$$\Psi^b_I = \max_{j \in \{1,2,\ldots,M-k\}} \{\Psi^{k+j,b}\}. \tag{D.2}$$

Intuitively, $\Psi^b_I$ measures the performance of the best fitting model that augments the pre-selected regression model with one variable from the list of orthogonalized candidate variables.

The max statistic controls for data snooping bias. With $M - k$ factors to choose from, the factor that is selected may appear to be significant by random chance. We adopt the max statistic as our test statistic to control for multiple hypothesis testing, similar to White (2000), Sullivan, Timmermann and White (1999) and FSW. Our bootstrap approach allows us to obtain the empirical distribution of the max statistic under the joint null hypothesis that none of the $M - k$ variables is true. Due to multiple testing, this distribution is very different from the null distribution of the test statistic in a single test. We make inference by comparing the realized (in the data) max statistic to this distribution.

Which statistic should we use to summarize the additional contribution of a variable in the candidate list? Depending on the regression model, the choice varies. For instance, in predictive regressions, we typically use the $R^2$ or the adjusted $R^2$ as the summary statistic. In cross-sectional regressions, we use the $t$-statistic to test whether

[3] The number of bootstrap samples ($B$) may impact the outcome of the test. While a larger $B$ usually results in a more accurate test, in reality test accuracy has to be traded off against computational intensity. Davidson and MacKinnon (2000) show that hundreds of samples usually suffice to achieve a reasonably accurate test. To be conservative, we set $B$ at 10,000 for our main results. For simulations, for which computational intensity is a concern, we set $B$ at 1,000.

the average slope is significant.[4] One appealing feature of our method is that it does not require an explicit expression for the null distribution of the test statistic. It therefore can easily accommodate different types of summary statistics. In contrast, FSW focuses only on the $R^2$ statistic and analytically derives the distribution of the max $R^2$ statistic.

For the rest of the description of our method, we assume that the statistic that measures the incremental contribution of a variable from the candidate list is given and generically denote it as $\Psi_I$ or $\Psi_I^b$ for the $b$-th bootstrapped sample.

We bootstrap $B = 10,000$ times to obtain the collection $\{\Psi_I^b, b = 1, 2, \ldots, B\}$, denoted as $(\Psi_I)^B$, i.e.,

$$(\Psi_I)^B = \{\Psi_I^b, b = 1, 2, \ldots, B\}. \tag{D.3}$$

This is the empirical distribution of $\Psi_I$ which measures the maximal additional contribution to the regression model when one of the orthogonalized regressors is considered. Given that none of these orthogonalized regressors is a true predictor in population, $(\Psi_I)^B$ gives the distribution for this maximal additional contribution when the null hypothesis is true, i.e., the null of no predictability of the $M - k$ candidate variables is true. $(\Psi_I)^B$ is the bootstrapped analogue of the distribution for maximal $R^2$'s in FSW. Similar to White (2000) and advantageous over FSW,[5] our bootstrap method is essentially distribution-free and provides a convenient and accurate approach to calculate the distribution of the test statistic through sample perturbations.[6]

Our bootstrap sample has the same number of time periods as the original data. This allows us to match the sampling uncertainty of the original data with the bootstrapped sample. When there is little time dependence in the data, we simply treat each time period as the sampling unit and sample with replacement. When time dependence is an issue, we use a block bootstrap, as explained in detail in the on-line appendix. In either case, we only resample the time periods. We keep the cross-section intact to preserve the contemporaneous dependence among the variables.

[4] In cross-sectional regressions, sometimes we use the average pricing errors (e.g., mean absolute pricing error) as the summary statistic. In this case, $\Psi^{eb}$ represents the minimum among the average pricing errors for the candidate variables.

[5] We are able to generalize FSW in two important ways. First, our approach allows us to maintain the distributional information among the regressors, helping us avoid the Bonferroni type of approximation in equation (3) of FSW. Second, even in the case of independence, our use of bootstrap takes the sampling uncertainty into account, providing a finite sample version of what is given in equation (2) of FSW.

[6] By treating the data as if they were the "population", the bootstrap allows us to obtain the exact distribution of the test statistic under this population with arbitrary accuracy by having a sufficiently large number of bootstrap iterations. It also yields an approximation to the distribution of the test statistic for the actual underlying population. Under regularity conditions, such an approximation is usually at least as accurate as the approximation obtained through first-order asymptotic theory (see, e.g., Horowitz et al. 2001).

***Step III: Hypothesis Testing and Variable Selection***

Working on the original data matrix $X$, we can obtain a $\Psi_I$ statistic that measures the maximal additional contribution of a candidate variable. We denote this statistic as $\Psi_I^d$. Hypothesis testing for the existence of the $(k+1)$-th significant predictor amounts to comparing $\Psi_I^d$ with the distribution of $\Psi_I$ under the null hypothesis, i.e., $(\Psi_I)^B$. With a pre-specified significance level of $\alpha$, say 5%, we reject the null if $\Psi_I^d$ exceeds the $(1-\alpha)$-th percentile of $(\Psi_I)^B$, that is,

$$\Psi_I^d > (\Psi_I)^B_{1-\alpha}, \tag{D.4}$$

where $(\Psi_I)^B_{1-\alpha}$ is the $(1-\alpha)$-th percentile of $(\Psi_I)^B$.

The result of the hypothesis test tells us whether there exists a significant predictor among the remaining $M-k$ candidate variables, after taking multiple testing into account. If the decision is positive, we declare the variable with the largest test statistic (i.e., $\Psi_I^d$) as significant and include it in a now augmented list of pre-selected variables. We then start over from Step I to test for the next predictor, if not all predictors have been selected. Otherwise, we terminate the algorithm and conclude that the pre-selected $k$ variables are the only ones that are significant.

## D.2 Fama-MacBeth Regressions

### D.2.1 Intuition

Our method can also be adapted to test factor models in cross-sectional regressions. We use the Fama-MacBeth regression (Fama and MacBeth, 1973) as an example.

One hurdle in applying our method to FM regressions is the time-varying slopes in cross-sectional regressions. In particular, separate cross-sectional regressions are performed for each time period to obtain a collection of cross-sectional regression slopes. We test the significance of a factor by looking at the time averaged cross-sectional slope coefficient. Therefore, in the FM framework, the null hypothesis is that the slope is zero in population. We adjust our method such that this condition exactly holds in-sample for the adjusted regressors.

First, we need to orthogonalize. Suppose we run a FM regression on a baseline model and obtain the panel of residual excess returns. In particular, at time $t$, let the vector of residual excess returns be $Y_t$. We are testing the incremental contribution of a candidate factor in explaining the cross-section of expected returns. Let the vector of risk loadings (i.e., $\beta$'s) for the candidate factor be $X_t$. Suppose there are $n_t$ assets in the cross-section at time $t$ so the dimension of both $Y_t$ and $X_t$ is $n_t \times 1$. Notice that $n_t$ can be time-dependent as it is straightforward for our method to handle unbalanced panels. In a typical FM regression, we would project $Y_t$ onto $X_t$. For our orthogonalization to work, we reverse the process, similar to what we do in predictive regressions. More specifically, we stack the collection of $Y_t$'s and $X_t$'s into two column vectors that have a dimension of $\sum_{t=1}^{T} n_t \times 1$, and run the following constrained regression model:

$$\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_T \end{bmatrix}_{\sum_{t=1}^{T} n_t \times 1} = \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_T \end{bmatrix}_{\sum_{t=1}^{T} n_t \times 1} + \xi_{1\times 1} \cdot \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_T \end{bmatrix}_{\sum_{t=1}^{T} n_t \times 1} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_T \end{bmatrix}_{\sum_{t=1}^{T} n_t \times 1}, \tag{D.5}$$

where $\phi_t$ is the constant vector of intercepts for time $t$, $\xi_{1\times 1}$ is a scalar, and $[\varepsilon_1', \varepsilon_2', \ldots, \varepsilon_T']'$ is the vector of projected regressors that will be used in the follow-up bootstrap analysis. This is a constrained regression as we have a single regression slope (i.e., $\xi$) throughout the sample. Had we allowed different slopes across time, we would have the usual unconstrained regression model where $X_t$ is projected onto $Y_t$ period-by-period. Having a single slope coefficient is key for us to achieve the null hypothesis in-sample for the FM model.

Alternatively, we can view the above regression model as an adaptation of the orthogonalization procedure that we use in predictive regressions. It pools returns and factor loadings together to estimate a single slope coefficient. What is different, however, is the use of separate intercepts for different time periods. This is natural since the FM procedure allows time-varying intercepts and slopes. To purge the variation in $Y_t$'s out of $X_t$'s, we need to allow for time-varying intercepts as well. Mathematically, the time-dependent intercepts allow the regression residuals to sum up to zero within each period. This property proves very important in that it allows us to form the FM null hypothesis in-sample, as we shall see later.

Next, we scale each residual vector $\varepsilon$ by its sum of squares $\varepsilon'\varepsilon$ and generate the orthogonalized regressor vectors:

$$X_t^e = \varepsilon_t/(\varepsilon_t'\varepsilon_t), \ t = 1, 2, \ldots, T. \tag{D.6}$$

These orthogonalized regressors are the FM counterparts of the orthogonalized regressors in predictive regressions. They satisfy the FM null hypothesis in cross-sectional regressions. In particular, suppose we run cross-sectional OLS with these orthogonalized regressor vectors for each period:

$$Y_t = \mu_t + \gamma_t X_t^e + \eta_t, \ t = 1, 2, \ldots, T, \tag{D.7}$$

where $\mu_t$ is the $n_t \times 1$ vector of intercepts, $\gamma_t$ is the scalar slope for the $t$-th period, and $\eta_t$ is the $n_t \times 1$ vector of residuals. We show in the next section that the following FM null hypothesis holds in-sample:

$$\sum_{t=1}^{T} \gamma_t = 0. \tag{D.8}$$

The above orthogonalization is the only step that we need to adapt to apply our method to the FM procedure. The rest of our method follows for factor selection in FM regressions. In particular, with a pre-selected set of right-hand side variables, we orthogonalize the rest of the right-hand side variables to form the joint null hypothesis that none of them is a true factor. We then bootstrap to test this null hypothesis. If we reject, we add the most significant one to the list of pre-selected variables and start over to test the next variable. Otherwise, we stop and end up with the set of pre-selected variables.

### D.2.2 Proof

The corresponding objective function for the regression model in (D.5) is given by:

$$\mathcal{L} = \sum_{t=1}^{T}[X_t - (\phi_t + \xi Y_t)]'[X_t - (\phi_t + \xi Y_t)]. \tag{D.9}$$

Taking first order derivatives with respect to $\{\phi_t\}_{t=1}^{T}$ and $\xi$, respectively, we have

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \phi_t} &= \iota_t' \varepsilon_t = 0,\ t = 1, \ldots, T, & \text{(D.10)} \\ \frac{\partial \mathcal{L}}{\partial \xi} &= \sum_{t=1}^{T} Y_t' \varepsilon_t = 0, & \text{(D.11)} \end{aligned}$$

where $\iota_t$ is a $n_t \times 1$ vector of ones. (D.10) says that the residuals within each time period sum up to zero, and (D.11) says that the $Y_t$'s are on average orthogonal to the $\varepsilon_t$'s across time. Importantly, $Y_t$ is not necessarily orthogonal to $\varepsilon_t$ within each time period. As explained in the main text, we next define the orthogonalized regressor $X_t^e$ as the rescaled residuals, i.e.,

$$X_t^e = \varepsilon_t/(\varepsilon_t'\varepsilon_t),\ t = 1, \ldots, T. \tag{D.12}$$

Solving the OLS (D.7) for each time period, we have:

$$\begin{aligned} \gamma_t &= (X_t^{e\prime} X_t^e)^{-1} X_t^{e\prime}(Y_t - \iota_t \mu_t), & \text{(D.13)} \\ &= (X_t^{e\prime} X_t^e)^{-1} X_t^{e\prime} Y_t - (X_t^{e\prime} X_t^e)^{-1} X_t^{e\prime} \iota_t \mu_t,\ t = 1, \ldots, T. & \text{(D.14)} \end{aligned}$$

We calculate the two components in (D.14) separately. First, notice $X_t^e$ is a rescaled version of $\varepsilon_t$. By (D.10), the second component (i.e., $(X_t^{e\prime} X_t^e)^{-1} X_t^{e\prime} \iota_t \mu_t$) equals zero. The first component is calculated as:

$$\begin{aligned} (X_t^{e\prime} X_t^e)^{-1} X_t^{e\prime} Y_t &= [(\frac{\varepsilon_t}{\varepsilon_t'\varepsilon_t})'(\frac{\varepsilon_t}{\varepsilon_t'\varepsilon_t})]^{-1}(\frac{\varepsilon_t}{\varepsilon_t'\varepsilon_t})' Y_t, & \text{(D.15)} \\ &= \varepsilon_t' Y_t,\ t = 1, \ldots, T, & \text{(D.16)} \end{aligned}$$

where we again use the definition of $X_t^e$ in equation (D.15). Hence, we have:

$$\gamma_t = \varepsilon_t' Y_t, \ t = 1, \ldots, T. \tag{D.17}$$

Finally, applying (D.11), we have:

$$\sum_{t=1}^{T} \gamma_t = \sum_{t=1}^{T} \varepsilon_t' Y_t = 0.$$

# E Model Discussion

## E.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 or forecasting variables so they appear irrelevant in-sample. That is, they achieve the null hypotheses in sample. However, the null varies across the regression models. As a result, a particular orthogonalization method that works in one model may not work in another model. For instance, in the panel regression model the null is that a factor does not help reduce the cross-section of pricing errors. In contrast, in Fama-MacBeth type of cross-sectional regressions, the null is that the time averaged slope coefficients is zero. Orthogonalizing factors as what we do in panel regressions, while keeping the time-series regression intercepts intact and thus helping achieve the null in panel regressions, will not achieve the desired null (i.e., a time averaged slope coefficient of zero) in the FM regressions.

Our method builds on the statistics literature on bootstraping. Jeong and Maddala (1993) suggest that there are two uses of the bootstrap approach that can be justified both theoretically and empirically. First, the bootstrap method provides a way to conduct statistical analysis (e.g., hypothesis tests, confidence intervals, etc.) when asymptotic theory is not tractable for certain models.[7] Second, even when asymptotic theory (or finite sample distribution under normality) is available, it may not be accurate in smaller samples.[8]

Our approach solves at least two problems. First, it is a daunting task to derive asymptotic distributions given the complicated structure of the cross-section of equity returns, e.g., unbalanced panel, cross-sectional dependency, number of firms is large relative to the number of time periods, etc. Second, as we mentioned previously, the GRS test is distorted when the returns for test portfolios are non-normally distributed. This problem is likely to be even worse given our use of individual stocks as test assets. Our bootstrap method allows us to overcome these difficulties and conduct robust statistical inference.

More specially, our method falls into the category of the nonparametric bootstrap that is routinely used for hypothesis testing. Hall and Wilson (1991) provide two valuable guidelines. The first, which can have a large impact on test power, is that bootstrap resampling should be done in a way that reflects the null hypothesis, even

[7] This is particularly relevant for our analysis that uses the max statistic. Important examples in the finance literature include Kosowski, Timmermann, Wermers, and White (2006) and Fama and French (2010).

[8] For other references on bootstrapping and its applications to financial time series, see Li and Maddala (1996), Veall (1992, 1998), Efron and Tibshirani (1993), and MacKinnon (2006). In our context, for example, Affleck-Graves and McDonald (1989) evaluate the finite sample performance of the GRS test under non-normality.

if the true hypothesis is distant from the null.[9] The second is to use pivotal statistics, that is, statistics whose distributions do not depend on unknown parameters.[10]

The design of our tests closely follows these principles. Take our panel regression model as an example. The first step orthogonalization, which is core to our method, ensures that the null hypothesis that a factor has no explanatory power for the cross-section of expected returns is exactly achieved in-sample. Our method therefore abides by the first principle and potentially has a higher test power compared to alternative designs of the hypothesis tests. In addition, when constructing the test statistics corresponding to the panel regression model, we make sure that pivotal statistics (e.g., $t$-statistics of the regression intercepts) are considered along with other test statistics.

## E.2 "Useful" Factors

To provide some context for our paper in the existing literature, we first clarify what we mean by claiming that a risk factor is "useful" in explaining the cross-section of expected returns. Using our notation for panel regressions from the previous session, let $\mathbf{a}$ and $\mathbf{a}_+$ be the vector of the regression intercepts for the baseline model and the augmented model ('+'). Our null hypothesis is:

$$H_0 : \mathbf{a}_+ = \mathbf{a},$$

against the alternative hypothesis,

$$H_a : \mathbf{a}_+ \neq \mathbf{a}.$$

Our method allows us to exactly achieve the null hypothesis in-sample by adjusting the regressors that are in the augmented model but are not in the baseline model. Under this hypothesis, we can simulate to find the distribution of any test statistic that involves the estimates of $\mathbf{a}_+$ and $\mathbf{a}$. The particular test statistics we examine later — consistent with our goal of identifying risk factors that help explain the cross-section of expected returns — are the ones that calculate the reduction in the averaged (equally weighted or value weighted) absolute intercept of the augmented

[9] Young (1986), Beran (1988) and Hinkley (1989) discuss the first guideline in more detail.

[10] To give an example of the use of pivotal statistics in bootstrap hypothesis testing, suppose our sample is $\{x_1, x_2, \ldots, x_n\}$ and the hypothesis under test is that the population mean equals $\theta_0$, i.e., $H_0 : \theta = \theta_0$. A test statistic one may want to use is $\hat{\theta}^* - \theta_0$, where $\hat{\theta}^* = \sum_{i=1}^{n} x_i/n$ is the sample mean. However, this statistic is not pivotal in that its distribution depends on the population standard deviation $\sigma$, which is an unknown parameter. According to Hall and Wilson (1991), a better statistic is to divide $\hat{\theta}^* - \theta_0$ by $\hat{\sigma}^*$, where $\hat{\sigma}^*$ is the standard deviation estimate. The new test statistic $(\hat{\theta}^* - \theta_0)/\hat{\sigma}^*$ is an example of a pivotal test statistic.

model relative to the baseline model. A positive reduction is regarded as evidence inconsistent with the null hypothesis.[11]

Suppose the augmented model has one additional factor relative to the baseline model. We consider it a "useful" risk factor (in addition to factors in the baseline model) if we reject the null hypothesis. More specifically, in the language of the expected return beta representation, a risk factor is considered useful if, relative to the baseline model, the inclusion of the risk factor in the baseline model helps reduce the magnitude of the cross-section of intercepts under the baseline model.

Our definition of a useful risk factor is consistent with Gibbons, Ross and Shanken (1989) under their assumption (or null hypothesis) that the augmented model is correctly specified, that is, $\mathbf{a}_+ = 0$. Under this assumption, the GRS approach would declare all factors in the augmented model necessary in explaining the cross-section of expected returns, hence useful risk factors. In our framework, given that we adopt a sequential selection procedure, all factors in the baseline model are already declared useful. Hence, $\mathbf{a}_+ = 0$ would imply that dropping the candidate factor from the augmented model will generate a non-zero intercept under the baseline model (i.e., $\mathbf{a} \neq 0$), which would be considered as evidence against the null hypothesis (i.e., it is better to include the candidate factor in the augmented model).[12] Hence, a rejection of the null hypothesis in our framework would identify the factor as useful, the same as in the GRS framework.

While the GRS approach is a model misspecification test and relies on the null hypothesis of $\mathbf{a}_+ = 0$, our test is anchored around the less ambitious null of $\mathbf{a}_+ = \mathbf{a}$, that is, the augmented model does not improve on the baseline model. Given the fact that hundreds of potential risk factors have been discovered and that we use individual stocks as test assets, it seems unlikely for one to stumble upon a correctly specified (augmented) model, that is, $\mathbf{a}_+ = 0$ for the cross-section of individual stocks.[13] Our approach is potentially more useful than GRS in that we do not require a correctly specified model, but instead look for any improvement the augmented model may have upon the baseline model.[14]

While our discussion above draws on the GRS framework to define a factor as useful from a risk perspective, recent papers by Stambaugh and Yuan (2016) and

[11] Notice that we are focusing on a one-sided test: only when there is a significant *positive* reduction in the averaged absolute intercept will we reject the null hypothesis. This is consistent with our goal of identifying risk factors that help explain the cross-section of expected returns.

[12] In the rare case where the factor is redundant in the sense that the pseudo factor (as defined in the previous section) is the same as the original factor, our method will declare the factor useless.

[13] Several issues make it difficult for us to find a correctly specified model for the cross-section of individual stocks. First, given the hundreds of factors discovered, it is currently infeasible to examine all the possible combinations of factors. Second, there might be omitted variables in that some useful risk factors may have not been discovered yet. Third, due to microstructure noise, certain stocks may not be fairly priced even if we have identified all the relevant risk factors.

[14] See Harvey (2017) for a related discussion of how improbable null hypotheses affect hypothesis testing.

Kozak, Nagel, and Santosh (2017) show that mispricing variables can also be disguised as “risk factors” to explain expected stock returns. Therefore, without taking a stand on the nature of the factor (i.e., either risk-based or mispricing-based), a useful factor in our framework should be interpreted as a factor whose factor loadings help explain the cross-section of stock returns based on our test statistics.

Finally, note that our discussion above is geared towards tests for which the left-hand assets do not necessarily include the right-hand factors, e.g., in our main application where we use individual stocks as test assets. When the left-hand assets span the right-hand factors, Barillas and Shanken (2017) provide a simple way to compare competing models. See our extensive discussion below for the relation between our approach and Barillas and Shanken (2017).

## E.3 Multiple Hypothesis Testing

Several extant papers also make attempts to control for multiple hypothesis testing in factor tests. Harvey, Liu, and Zhu (2016) explore several statistical methods that aim to control the FWER or alternative definitions of error rates (e.g., the false discovery rate, which is the expectation of the fraction of false discoveries among all discoveries, denoted as FDR). Feng, Giglio, and Xiu (2019) propose a two-stage LASSO approach that asymptotically selects the correct model with probability one. In contrast to these papers, our reliance on White (2000)’s bootstrap approach allows us to control the FWER in finite samples.[15] We believe the finite-sample perspective of our method is particularly attractive given the number of factors discovered is likely a function of T (the number of time periods, see the evidence in Harvey, Liu, and Zhu, 2016, and Harvey and Liu, 2019), which may complicate the asymptotic analysis of LASSO-based approaches. In addition, similar to the goal of White (2000), our bootstrap-based approach is straightforward to use and applicable to a variety of situations on factor tests. Successful applications of White (2000)’s idea in finance research includes Sullivan, Timmermann, and White (1999), which control for data-snooping bias in testing technical trading rules, and Fama and French (2010), which evaluate mutual fund performance by controlling for a large number of tests in the cross section. Tests of factor models face similar challenges to these two areas of research given the concern about data snooping bias.

## E.4 Sequential Tests

Our framework allows one to sequentially build up the factor model. The sequential nature of our test implies that the final factor set we arrive at is path dependent and may not converge to the underlying true model with a probability of one, if such

[15] This is true up to the precision offered by the bootstrap approximation to the true data generating process.

a model exists. This is well-known issue in statistics and the typical solution is to invoke model selection consistent criteria such as the Akaike information criterion (AIC) or the Bayesian information criterion (BIC). In the context of the literature on factor tests, Kogan and Tian (2017) explore permutation tests for all combinations of five-factor models among a given set of factors and find that the five-factor model in Fama and French (2017) is far from being the optimal combination when compared to alternative five-factor models. Whereas Kogan and Tian (2017) constrain themselves to five-factor models, alternative ways of searching through the entire model space have been proposed by Barillas and Shanken (2018) and Chib, Zeng, and Zhao (2019).

Bearing in mind the potential selection inconsistency of our approach, one advantage of having a sequential model is to have a clear ordering of the importance of the variables being selected at each stage of our tests. This may make it easier to interpret the test outcome from an economic standpoint. For example, suppose two factors, when combined together, subsume the explanatory power of the market factor. As a result, the market factor may never be selected for the final model when all combinations of factors are exhausted. In our framework, however, if the market factor dominates each of the other two factors and survives the multiple testing threshold, it will be selected for the final model. This is indeed what we find in our empirical analysis. Our approach is thus consistent with the practice in economic research where pre-established results are weighed heavily and controlled for in subsequent research, regardless of whether pre-established results survive the search of the best performing model when all results enter on an equal footing. In general, the convenience and interpretability of sequential methods prompts the model selection literature in statistics to bifurcate into two areas of research: one is the sequential approach that allows one to build up increasingly complex models,[16] and the other is the theoretically more appealing approach that aims to select the true model with a probability converging to one. From this perspective, the recent development of factor tests mirrors this bifurcation: whereas our paper sequentially builds up the factor model, the aforementioned papers seek to find the best combination of factors. We therefore view our approach as complementing Kogan and Tian (2017), Barillas and Shanken (2018) and Chib, Zeng, and Zhao (2019).

We also want to point out that whether or not our approach is sequential does not affect the validity of our approach 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 pre-existing factors)?

Having said this, when two variables are close in performance at a certain stage of our test, whether or not one variable is selected versus the other does raise the concern about its potential impact on the selection of additional factors. In such cases, we experiment with selecting the second best and examine its impact on the subsequent selection. We illustrate how this can be done with examples in our empirical analysis.

[16] Examples in the regression context include forward stepwise linear regression, least angle regression, and the ever-active path in LASSO regression.

## E.5 Related Literature

**The GRS Test** Our first contribution is to extend the insights of Gibbons, Ross and Shanken (GRS, 1989) by proposing a new framework that is applicable to a large collection of assets. Our method can be applied to portfolios or individual assets. For a long time, the literature has struggled with defining the appropriate test portfolios. Indeed, it is confusing to find that, for example, size and book-to-market "work" on portfolios sorted by size and book-to-market but these same factors do not "work" when under a different portfolio sorting variable. Indeed, there are hundreds (if not thousands) of different ways to form portfolios.

We advocate the use of individual assets which leads to a much larger dimension of the cross-section than the time-series. While a large cross-section of assets present a challenge to the GRS statistic, our framework allows for a larger cross-section than time-series by sidestepping the estimation of the large dimensional residual covariance matrix in the GRS statistic. Importantly, we provide bootstrap-based inference to our test statistics, which no longer follow a standard $F$-distribution (as for the GRS statistic) when we use alternative weighting schemes to weight the cross-section of pricing errors.

We are aware that our testing framework may have ambiguous implications for certain assets. For example, suppose the baseline model is a constant and the augmented model is the CAPM. Suppose for a certain stock A, its excess return is 10% per annum, which would be its regression intercept under the baseline model. To test the market factor, our approach looks at the difference in absolute intercept between CAPM and the baseline model. A significant reduction of the absolute intercept by CAPM would be regarded as evidence against the null model (i.e., a constant). Notice that under the GRS null hypothesis that $\mathbf{a}_{+} = 0$ (i.e., the market factor is mean-variance efficient), the sign of our test statistic would be unambiguous for stock A: the 10% premium under the baseline model should be completely absorbed by the market factor under CAPM, suggesting a 10 percentage point reduction in absolute intercept. When $\mathbf{a}_{+} = 0$ is not true for CAPM, it is possible that asset A may have an increase in absolute intercept with the inclusion of the market factor even if the market factor is a useful risk factor (e.g., the market factor is one of the factors for the underlying multi-factor model). The reason is that there are other factors in the underlying multi-factor model in addition to the market factor that also carry a premium. In essence, this is a omitted variable problem that affects not only our model, but also most asset pricing tests, including both cross-sectional and time-series tests.[17] We believe that our tests are more robust to omitted variables than several existing methods in that it is more plausible to think that the augmented model (i.e., CAPM in our example) helps reduce the absolute intercepts for the majority of assets than that the augmented model achieves $\mathbf{a}_{+} = 0$ for the cross-section of individual stocks for both the GRS approach and the cross-sectional approach.

[17] See, e.g., Jagannathan and Wang (1998), Shanken and Zhou (2007), Kan and Robotti (2008, 2009), Kleibergen (2009), Kleibergen and Zhan (2014), and Giglio and Xiu (2019).

Having defined what we mean by a useful factor, we discuss two related issues in greater detail. First, what do we use as test assets and why would they matter to the identification of useful risk factors? Second, what test statistics do we use to quantify the reduction in alphas? Before we elaborate on these issues, let us again revisit the GRS approach.

GRS can be used to test whether a certain set of factors is mean-variance efficient or, as recently advocated by Barillas and Shanken (2017), to compare competing factor models under certain circumstances. We first focus on the original use of GRS as a test of the mean-variance efficiency of a factor portfolio.

Given a set of test assets $A$ and a candidate model $G$, GRS can be interpreted as the difference between $SR^m(A, G)$ and $SR^m(G)$, where $SR^m(A, G)$ denotes the maximum ex post Sharpe-ratio constructed using both $A$ and $G$ and $SR^m(G)$ denotes the maximum ex post Sharpe-ratio using $G$ only (see GRS, Barillas and Shanken, 2017, Fama and French, 2015b, 2016). Hence, GRS measures the additional contribution of $A$ relative to $G$ in generating a more efficient mean-variance frontier. While the GRS approach is useful and intuitive when the cross-sectional dimension is much smaller than the time-series dimension (i.e., $N \ll T$), its use is limited when the cross-sectional dimension is close to or even larger than the time-series dimension, both from a statistical perspective and an economic perspective. Statistically, when we have more assets than the number of time-series observations, the covariance matrix is not of full rank, so its inverse — which is required for the calculation of the GRS statistic — does not exist.[18] More importantly, the statistical issue with the GRS statistic alludes to a bigger problem from an economic perspective: when there are more assets than the number of observations, there always exist arbitrage opportunities in-sample, implying an infinitely large $SR^m(A, G)$ ex post. Hence, tests that are based on the ex post efficiency of asset returns, such as the GRS statistic, are not appropriate for applications where the cross-sectional dimension is large.

While a large cross-section of assets present a challenge for the GRS statistic, recent advances in asset pricing research call for methods that can deal with a large cross-section of assets. Harvey, Liu and Zhu (2016) document hundreds of anomalies discovered by the literature. Hence, a comprehensive evaluation of a factor model would include at least hundreds of test assets, if not more. Hou et al. (2015, 2016) confront new generations of factor models with hundreds of anomalies, one at a time. However, one would expect a joint test in the spirit of GRS to be more powerful in disentangling candidate models. Indeed, as GRS illustrate, a joint test can produce different test results than a set of univariate tests. Finally, as shown in GRS and MacKinlay (1987), the manner in which test assets are grouped into portfolios can affect the value of the test statistic and the conclusion of the test. In this paper, we advocate the use of individual stocks to avoid these issues with portfolio formation.

[18] See Bai and Shi (2011) for a recent survey on the estimation of the covariance matrix for finance applications.

Given the difficulty of the GRS in coping with a large cross-section of assets, we propose a new set of test statistics and provide a flexible approach to make inference. In essence, any test of a factor model involves the evaluation of the profitability of strategies that provide risk adjusted returns beyond what the benchmark risk factors produce. In the case of the GRS approach, this profitability is given by $\alpha'\Sigma^{-1}\alpha$, where $\alpha$ is the cross-section of alphas and $\Sigma$ is the residual covariance matrix. However, several issues exist with the use of this metric to evaluate candidate models when we have a large cross-section of assets. As emphasized by Fama and French (2015b), ex post estimation of $\Sigma^{-1}$ based on portfolios ($N \ll T$ still holds) leads to implausible magnitudes of short positions in certain portfolios, casting doubt on the economic interpretation of $\alpha'\Sigma^{-1}\alpha$. When $N$ becomes even larger compared to $T$, as we discussed previously, $\alpha'\Sigma^{-1}\alpha$ may be infinitely large. These issues are well-known to the literature on portfolio optimization.[19] Ex post optimized portfolios provide upwardly biased estimates of the actual performance of the portfolio that is ex ante efficient. This bias is potentially very large for our applications. Then, the question is: if the ex post optimized portfolio (in the mean-variance sense) cannot truly represent the profitability of strategies that are not accounted for by their exposures to the benchmark risk factors, what portfolios can?

We provide several test statistics that are both economically meaningful and statistically sound, while at the same time avoiding the drawbacks of the GRS approach when applied to a large number of test assets.[20]

The general idea behind our test statistics is to examine the weighted average of the cross-section of absolute alphas, with the weights pre-determined and capturing economically motivated moments of alphas.[21] We explore three specifications for weights: equal weighting, weights being the residual standard deviations from a factor model regression, and market capitalization weighted residual standard deviations. Compared to the weighting scheme in the GRS approach, our method pays more attention to the first moment of returns than the second moments, in particular the cross correlations among the assets. We believe that this is appropriate in our context

[19] See, e.g., Green and Hollifield (1992), Ledoit and Wolf (2003), Jagannathan and Ma (2003), and DeMiguel, Garlappi, Nogales, and Uppal (2009).

[20] Other papers that extend the idea of the GRS approach include Shanken (1990) and MacKinlay and Richardson (1991). Shanken (1990) examines the conditional efficiency of benchmark factors using size and industry portfolios. He pays particular attention to time-varying betas and conditional heteroskedasticity by using observable state variables as instruments to model betas and residual variances. MacKinlay and Richardson (1991) develop tests of unconditional mean-variance efficiency using the GMM (Generalized Method of Moments) framework, which is robust to nonnormality and heteroskedasticity in asset returns. Different from these papers, we focus on test statistics that can be applied to a large cross-section of assets. Our test statistics also control for cross-sectional dependency and data snooping bias, and allow more flexible weighting schemes of the cross-section of alphas. These features are difficult to achieve within the frameworks of the above papers that explore the asymptotic variations of the GRS approach.

[21] Papers that also examine pricing errors include, e.g., Shanken (1987), Ferson and Harvey (1991), Geweke and Zhou (1996), and Fama and French (2015a). Different from these papers, our bootstrap-based method allows us to make statistical inference on moments of the cross-section of pricing errors.

since the primary concern for the research on factor models is to explain patterns in the cross-section of expected returns. This is also consistent with the goal of most investment strategies in practice that focus on the cross-section of alphas. Finally, our bootstrap-based approach is flexible in providing inference for sophisticated weighting schemes such as value weighting. This is difficult to achieve for traditional methods that rely on asymptotic approximations. Given the importance of weighting schemes for asset pricing tests in general (see, e.g., Cochrane, 2005, Ludvigson, 2011), we consider the flexibility of our framework in dealing with alternative weighting schemes crucial in assessing the all-around performance of a factor model.

Our test statistics are economically motivated. The arbitrage pricing theory (APT) of Ross (1976) shows that one of the important conditions for the expected return beta pricing representation to hold (relative to a factor model) is the absence of arbitrage opportunities in an efficient market. Hence, the extent to which arbitrage opportunities are available is an indication of the performance of a factor model. Our test statistics focus on alphas, which measure the mispricing of a security relative to a synthetic asset that is fairly priced, and are thus directly linked to the amount of arbitrage opportunities in the market.[22] Importantly, rather than trying to find the ex post efficient portfolio that weights the cross-section of alphas by the inverse of the residual covariance matrix as in the GRS statistic, we only use the diagonal elements of the residual covariance matrix or the value-weighted diagonal elements to weight the cross-section of alphas. This helps circumvent the issue of reliably estimating the inverse of the residual covariance matrix when the number of assets is close to or larger than the number of time-series observations. From an investment perspective, while the GRS approach tells us how to achieve the ex post efficient frontier, it becomes difficult for investors to follow the asset allocation rules dictated by the GRS approach when the number of assets is large since 1) the GRS statistic implies implausibly large short positions on certain assets even when $N$ is moderately large; 2) in-sample optimization leads to overfitting, making the in-sample allocation rules under the GRS approach suboptimal in the out-of-sample.[23] Our approach relies on simple trading strategies that are more accessible to an average investor that seeks to exploit (statistical) arbitrage opportunities.[24]

[22] Arbitrage in our context refers to alpha arbitrage, that is, exploiting the cross-section of mispricing (alphas) through statistical arbitrage.

[23] See, e.g., Hansen (2009), Ledoit and Wolf (2014), and Paulsen and Söhl (2016).

[24] Notice that unless an explicit out-of-sample forecasting exercise is performed, all tests are in-sample in nature and thus not "accessible" to investors from an ex ante investment perspective. However, if the data were stationary, we would expect the in-sample performance of investment strategies constructed based on in-sample test statistics to persist in the out-of-sample to some degree. The problem with an approach like the GRS that explores the in-sample optimized portfolio is that, at least for our application to a large number of individual stocks, in-sample optimization leads to overfitting and extreme positions in certain stocks, implying potentially a large degree of deterioration in performance of the optimized portfolio. In contrast, simple investment strategies that explore the average alpha in the cross-section are likely to experience a smaller degree of deterioration in performance. Existing papers on portfolio optimization have shown considerable evidence on the severe deterioration of in-sample optimized portfolios. For example, DeMiguel, Garlappi, and Uppal (2009) show that in-sample minimum variance portfolios are consistently outperformed

Our test statistics are also related to several existing papers on asset pricing tests. Geweke and Zhou (1996) use the averaged squared pricing errors from an APT as the model performance metric and derive its posterior distribution in a Bayesian framework.[25] Shanken (1987) derives testable restrictions on the pricing deviation for each individual asset. Hansen and Jagannathan (1997) and Kan and Robotti (2009) examine a quadratic form of the cross-section of pricing errors within a GMM framework. Fama and French (2015a) use similar test statistics to provide informal statistical inference using characteristics-sorted portfolios. Different from these papers, our focus in this paper is to provide flexible and rigorous inference on test statistics that capture the average pricing error for a large cross-section of test assets.

Our framework also builds on important insights from the literature on portfolio choice. DeMiguel, Garlappi, and Uppal (2009) show that naive portfolio strategies often dominate strategies that are based on in-sample mean-variance optimization when taken out-of-sample. Given the difficulty in reliably estimating the residual covariance matrix, various papers propose shrinkage type of estimators (see, e.g., Ledoit and Wolf, 2003, 2004, DeMiguel et al. 2009, Jagannathan and Ma, 2003, Elton, Gruber, and Spitzer, 2006). As we discussed previously, the evaluation of asset pricing models go hand in hand with the portfolio choice problem. To the extent that investors face parameter uncertainty when there are a large number of assets, simple trading strategies may be more relevant to investors than strategies implied by an in-sample mean-variance optimizer. As a result, tests based on simple trading strategies are likely more informative from an economic standpoint than those based on in-sample optimized strategies.

Our test statistics are also powerful from a statistical perspective. We are well aware of the fact that in theory, the GRS statistic is the most powerful test statistic in rejecting a factor model when its assumptions are satisfied. However, in practice, as shown in Affleck-Graves and McDonald (1990), the contribution of the off-diagonal elements of the residual covariance matrix is often limited in correctly rejecting a model. In fact, even with a modest number of assets, they show that the power of the GRS statistic is often higher when we only use the diagonal elements of the residual covariance matrix. The issue is likely to be even worse when $N$ approaches $T$ and in our application $N > T$. We therefore follow Affleck-Graves and McDonald and focus on the diagonal elements. Nonetheless, we pay particular attention to the power issue of our test statistics and provide simulation evidence on the power of our tests.

**Reconciling with Barillas and Shanken (2017)** Now returning to the questions raised previously, why do test assets matter in our framework? Let us consider an example. Suppose we try to compare the performance of two factor models: $G_1$ and

in out-of-sample tests by simply weighted portfolios, even with dozens of well-diversified industry or characteristic-based portfolios. We would expect the deterioration in performance of in-sample optimized portfolios to be even worse for our application with a large number of individual stocks.

[25] Harvey and Zhou (1990) use the full covariance matrix in a similar Bayesian framework as they only have a dozen test assets.

$G_2$. $G_1$ is a one-factor model that includes the market factor. $G_2$ is a two-factor model that includes the market factor and a long-short portfolio that takes positions in two stocks with extreme returns: one stock generates the highest in-sample mean return and the other generates the lowest in-sample mean return. To compare $G_1$ and $G_2$, we will use two sets of test assets. One set ($S_1$) only includes three assets: the market portfolio and the two stocks that have extreme returns in-sample. The other ($S_2$) includes the entire cross-section of stocks based on which the market factor is generated. Notice that the factors in $G_1$ and $G_2$ are spanned by either $S_1$ or $S_2$, satisfying the spanning condition in Barillas and Shanken (2017). Do we expect to see a different relative performance of model $G_1$ compared to model $G_2$ by using $S_1$ as test assets versus $S_2$?

Barillas and Shanken (2017) argue that there should be no difference in relative model performance in terms of the GRS statistic as long as the factors in the candidate models are spanned by test assets. The reason is, given that the GRS statistic can be (loosely) interpreted as the difference between the maximum ex post Sharpe ratio constructed using both test assets and the factors (which is the same as the maximum ex post Sharpe ratio constructed using test assets alone since factors are assumed to be spanned by test assets) and the maximum ex post Sharpe ratio constructed using only the factors,[26] the relative performance of $G_1$ versus $G_2$ is given by the difference in the maximum ex post Sharpe ratio between $G_1$ and $G_2$. In the example, since both $G_1$ and $G_2$ can be spanned by either $S_1$ or $S_2$, the relative performance of $G_1$ versus $G_2$ should be unchanged whether we use $S_1$ or $S_2$ as test assets. Given that $G_2$ includes the long-short portfolio that performs extremely well in-sample, the maximum ex post Sharpe ratio of $G_2$ will likely dominate that of $G_1$. Hence, $G_2$ appears to be a better model than $G_1$ whether we use $S_1$ or $S_2$ as test assets.

For our application, we believe the argument in Barillas and Shanken is problematic for the following reasons. Since we are looking at the cross-section of individual stocks, $N$ is large relative to $T$. This implies that $SR^m(S_2)$ (that maximum ex post Sharpe ratio for assets in $S_2$) will be much larger than either $SR^m(G_1)$ or $SR^m(G_2)$. For the sake of our argument, suppose $SR^m(S_2) = \infty$, which holds if $N > T$. Then $GRS(S_2, G_1) = SR^m(S_2) - SR^m(G_1) = \infty - SR^m(G_1) = \infty$ and $GRS(S_2, G_2) = SR^m(S_2) - SR^m(G_2) = \infty - SR^m(G_2) = \infty$.[27] Hence, since the GRS statistic under both $G_1$ and $G_2$ equals infinity, $G_2$ is no better than $G_1$, contrary to the above argument in Barillas and Shanken (2017). In essence, when both models imply large values for the GRS statistic (which will likely happen when we have a large cross-section of assets), the exercise of determining which model is less rejected in the data is not that meaningful.

[26] Technically speaking, the GRS statistic adjusts this difference by a factor that captures estimation uncertainty in factor betas.

[27] We assume that $SR^m(S_2) = \infty$ to better illustrate the intuition. All we need is for $SR^m(S_2)$ to be much larger than both $SR^m(G_1)$ and $SR^m(G_2)$, which is likely to be the case if we have a large number of assets in $S_2$. However, we do not require $N > T$.

Our framework uses a different test statistic than the one proposed in GRS and thus provides a different perspective on the use of test assets than Barillas and Shanken (2017). First, instead of using the ex post efficient portfolio, we use the weighted cross-section of absolute alphas as the test statistic, with the weights being the residual standard deviations or the value-weighted residual standard deviations. Notice that while the maximum Sharpe ratio of the ex post efficient portfolio may as well be infinity (e.g., when $N > T$), our test statistic is always well-defined and captures mispricing relative to a factor model. Second, we use the ratio statistic — the percentage change in the weighted cross-section of absolute alphas of the augmented model (i.e., $G_2$) relative to that of the baseline model (i.e., $G_1$) — instead of the difference test statistic in Barillas and Shanken (2017) to compare relative model performance. This ensures that significant improvement of the augmented model relative to the baseline model is likely to represent economically meaningful reduction in the weighted cross-section of absolute alphas. Indeed, relating to the previous paragraph, while $G_2$ dominates $G_1$ in terms of the in-sample maximum Sharpe ratio (which will be picked up by the difference test statistic), it offers little help in explaining the bulk of the variation in the cross-section of expected returns for individual stocks. A ratio-based statistic will likely correctly identify $G_2$ as insignificant.

In our framework, the relative performance of $G_1$ and $G_2$ will depend on the test assets we use. When $S_1$ is used, since it includes the two extreme stocks that are used to construct the long-short portfolio in $G_2$, the percentage reduction in the weighted cross-section of absolute alphas under $G_2$ is likely to be large. Hence, $G_2$ is likely to be declared the better model. This is analogous to the situation when we use the Fama-French factors to explain the returns of Fama-French portfolios (which are sorted on the same information as the factors) — Fama-French factors are likely to be significant. In contrast, when $S_2$ is used, since the majority of stocks in the cross-section are unlikely to be related to this ad-hoc long-short portfolio based on stocks with extreme returns (with the inclusion of the baseline factors in $G_1$, that is, the market factor), the percentage reduction in the weighted cross-section of absolute alphas under $G_2$ is likely to be small and statistically insignificant, leading us to reject $G_2$ as a better model. We believe that our framework leads to a more intuitive conclusion for our example than Barillas and Shanken (2017): it is highly unlikely that an ad-hoc long-short portfolio would substantially complement the market factor in explaining the returns of the cross-section of individual stocks, although it generates a large Sharpe ratio and thus appears to be able to generate a mean-variance more efficient portfolio when combined with the market factor.[28]

[28] Our argument is not specific to individual stocks. For example, suppose we have two factors. One ($f_1$) is Berkshire Hathaway's return minus the riskfree rate (or the extreme long-short portfolio in our example). The other ($f_2$) is the value strategy (i.e., the Fama-French HML factor). Test assets are the union of $f_1$ and the ten book-to-market sorted decile portfolios. Based on Barillas and Shanken (2017), $f_1$ would highly likely be considered as a better factor than $f_2$, given the extraordinary track record of Berkshire Hathaway. Berkshire Hathaway as a factor, however, lacks economic foundation and would unlikely be selected in our framework since $f_1$ is unlikely to be related to a large cross-section of individual stocks.

The implication of our framework that testing outcomes depend on test assets is not new to the literature. Pricing errors at the individual firm level may be distorted at the portfolio level, potentially leading to different test results using portfolios vs. stocks or using different portfolios (Roll, 1977, Kandel and Stambaugh, 1995, Ahn, Conrad, and Dittmar, 2003, Fama and French, 2008, Hoberg and Welch, 2009, Cederburg and O'Doherty, 2015). In addition, there is potentially a loss in efficiency in using a few selected portfolios rather than a broad cross-section of stocks (Litzenberger and Ramaswamy, 1979, Ang et al., 2016). Finally, the recent debate on the relative performance of the new generations of factor models (Hou et al., 2015, 2016, Fama and French, 2015a, 2018) highlights the important role of test assets in determining the testing outcomes.

**A SDF Interpretation** We use the stochastic discount factor (SDF) approach to offer an alternative way to interpret what we are doing. Let $A$ be the collection of test assets and suppose it spans the factors in model $G$. Suppose the true underlying SDF is $M$. Let the projection of $M$ on $A$ be $M_a = proj(M|A)$ and let the projection of $M$ on $G$ be $M_g = proj(M|G)$. Assuming $M_g$ is uniquely determined and given that $G$ is spanned by $A$, we must have $M_g = proj(proj(M|A)|G) = proj(M_a|G)$. If $M_g = M_a$, then the mean-variance frontier constructed with $G$ will be identical to the mean-variance frontier constructed with $A$, in which case $G$ should be declared as a successful factor model in explaining the returns for assets in $A$. However, in general $M_g \neq M_a$ and the discrepancy between $M_g$ and $M_a$ is an indication of the performance of $G$ is in pricing assets in $A$. Hansen and Jagannathan (1997) show that evaluating the distance between $M_g$ and $M_a$ in the space of pricing kernels is equivalent to the evaluation of the maximum distance between prices of a portfolio's payoff under the true model and the proposed model $G$. For our application, however, the maximum distance is likely to be very large and will achieve infinity if $N > T$.[29] Moreover, the asymptotic approximations that provide inference on the Hansen-Jagannathan distance are not applicable when $N > T$. In essence, these are attributable to the fact that the payoff space that includes all the combinations of the returns of assets in $A$ is too large, to the extent that arbitrage opportunities appear to exist in-sample when $N > T$.

Our approach, building on the insights of Hansen and Jagannathan (1997), also tries to evaluate model $G$ by examining the discrepancy in security price between the true model and the proposed model. However, instead of searching for the maximum price discrepancy within a large payoff space — so large that the maximum discrepancy approaches infinity, we pre-select a set of strategies that exploit mispricing in the market and test the statistical significance of the profitability of these strategies. Our approach is consistent with Fama and French (2015b)'s argument that the evaluation

[29] To see this, when $N > T$, there exist (in-sample) arbitrage opportunities. It is easy to combine an arbitrage strategy with a portfolio that generates a random payoff into a composite portfolio that generates a payoff distribution that has a unit $L_2$-norm but has an arbitrarily large price. Hence, the Hansen-Jagannathan distance is infinity.

of asset pricing models should be based on portfolio strategies that are economically meaningful. On the other hand, searching for the maximum price discrepancy as in Hansen and Jagannathan (1997) for a large set of assets may lead to "error maximization",[30] that is, the maximum price discrepancy is picking up unimportant features of the data due to estimation errors rather than measuring price discrepancies that can actually be exploited by (statistical) arbitragers. Our approach, which intentionally focuses on a set of alpha strategies that exploit mispricing, helps alleviate the concern of error maximization.

Kan and Robotti (KR, 2009) build on Hansen and Jagannathan (1997) and evaluate the difference of the Hansen-Jagannathan distance between competing models. Since we also look at the incremental contribution of a factor model, KR is related to our paper. However, KR and our paper differ in several dimensions. First, KR rely on the asymptotic distribution of the Hansen-Jagannathan distance when N is small and T is large, which diverges from our application. Second, even for $N < T$, our bootstrap-based approach allows us to control for cross-sectional dependence (possibly nonparametric) and non-normality in the data, which may result in finite sample (i.e., in $T$) distributions of the test statistics that are substantially different from those derived under asymptotic approximations. Third, while KR analytically derive the asymptotic distribution of their test statistic between a baseline model and an augmented model, it is not clear how to derive this distribution for the max test statistic when there are multiple competing augmented models. We achieve this and thus control for data snooping bias through the bootstrap. Finally, while the Hansen-Jagannathan distance weights pricing errors by a time-invariant matrix (i.e., an identity matrix or the inverse of the unconditional covariance matrix of returns), more interesting weighting schemes from an economic standpoint include value weighting, which is time-varying. Our approach is flexible in dealing with more sophisticated weighting schemes.

Another related approach is Lewellen, Nagel, and Shanken (LNS, 2010), who advocate the use of confidence intervals as opposed to traditional $p$-values for several test statistics. This avoids the assumption of zero pricing errors for the cross-section of assets for the traditional $p$-value-based approaches, and is consistent with our paper in trying to evaluate the incremental contribution a factor without assuming a correctly specified model. However, there are important differences between the two papers. First, while the main point of LNS is that the use of test assets with a strong factor structure may lead to spurious factors, we demonstrate a similar point by developing methods that can be applied to a large cross-section of assets that are not influenced by any particular characteristic sorts. For example, among the several test statistics examined by LNS, the one that is closest to ours is Shanken (1985)'s $T^2$ test statistic, which is akin to the GRS approach in that it evaluates the difference between the maximum generalized squared Sharpe ratio based on test assets and that attainable from the factors. Different from the $T^2$ test statistic, our paper focuses on pricing errors, driven by both statistical and economic concerns for our application to

[30] See Michaud (1989).

a large cross-section of test assets, as we discussed previously.[31] Second, we believe that one important mechanism through which spurious factors can be generated is data mining. As such, our method explicitly controls for test multiplicity using the bootstrap and is relevant to several fields in finance where data snooping bias is a serious concern.

Finally, in the presence of data mining, an issue prominent in the recent literature on the cross-section of expected stock returns, our approach has a key advantage over the GRS method. In particular, given two nested models ($M$: baseline model; $M_{aug}$: augmented model), it is not clear how one can make statistical inference to tell apart these models based on the GRS statistic. As we discussed previously, the issue is that under the GRS framework, the null of the augmented model $M_{aug}$ is not the same as $M$, that is, $M$ is the underlying factor model that is mean-variance efficient. Hence, we are not able to provide inference on the incremental contribution of $M_{aug}$ to $M$. In contrast, our test statistics set $M$ as the null model and evaluate the incremental contribution of $M_{aug}$. Importantly, our setup allows us to evaluate the significance of the maximal contribution of $M_{aug}$ when multiple $M_{aug}$'s have been tried (with the same baseline model $M$), thus controlling for multiple tests. Thus, our formulation bridges the literature on the cross-section of expected returns (for which we know factor dredging is a serious concern) and the literature on the bootstrap reality check (which provides a flexible and robust framework to control for data snooping). We show how the basic idea of our framework applies to a variety of regression setups that are commonly used in financial economics.

# F Results Based on Sorted Portfolios

## F.1 Fama-French 25 Portfolios

[31] Other test statistics examined by LNS include the cross-sectional $R^2$ and the generalized cross-sectional $R^2$, which are related to cross-sectional regressions such as the Fama-MacBeth approach. We show how to adapt the idea of our paper to Fama-MacBeth regressions in the appendix. Thus, we also provide one way to make robust inference on cross-sectional $R^2$'s.

Table F.1: **Summary Statistics on Fama-French 25 Portfolios, January 1968 - December 2012**

Summary statistics on Fama-French 25 portfolios. We report the mean annual excess returns for Fama-French size and book-to-market sorted 25 portfolios.

| | Low | 2 | 3 | 4 | High |
|---|---|---|---|---|---|
| Small | 0.009 | 0.078 | 0.085 | 0.106 | 0.120 |
| 2 | 0.039 | 0.074 | 0.095 | 0.101 | 0.108 |
| 3 | 0.047 | 0.082 | 0.082 | 0.093 | 0.119 |
| 4 | 0.062 | 0.061 | 0.077 | 0.087 | 0.090 |
| Big | 0.046 | 0.061 | 0.053 | 0.059 | 0.069 |

Table F.2: **Summary Statistics on Ten Market Beta-Sorted Portfolios, January 1968 - December 2012**

Summary statistics on ten market beta-sorted portfolios. We sort the cross-sectional of stocks each month into ten beta portfolios based on a rolling five-year market model regression. We report the mean annual excess returns for these portfolios. The sample period is from January 1968 to December 2012.

| Low | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | High |
|---|---|---|---|---|---|---|---|---|---|
| 0.009 | 0.078 | 0.085 | 0.106 | 0.120 | 0.009 | 0.078 | 0.085 | 0.106 | 0.120 |

Table F.3: **Summary Statistics on 90 Low-Turnover Anomaly Portfolios, January 1968 - December 2012**

Summary statistics on 90 anomaly portfolios. For the nine low-turnover anomalies in Novy-Marx and Velikov (2016), we report the mean annual excess returns from January 1968 to December 2012. "GP" is gross profitability; "Val" is value; "ValProf" is the combination of value and gross profitability calculated as the sum of the univariate ranks for book-to-market and profitability; "AssGr" is asset growth; "Inv" is investment; "Fscore" is Piotroski's F-score; and "NI" is net issuance.

| | Size | GP | Val | ValProf | Accruals | AssGr | Inv | Fscore | NI |
|---|---|---|---|---|---|---|---|---|---|
| Low | 0.106 | 0.053 | 0.020 | 0.011 | 0.131 | 0.159 | 0.152 | 0.077 | 0.132 |
| 2 | 0.061 | 0.080 | 0.055 | 0.049 | 0.120 | 0.132 | 0.130 | 0.075 | 0.102 |
| 3 | 0.075 | 0.073 | 0.072 | 0.061 | 0.124 | 0.117 | 0.123 | 0.099 | 0.098 |
| 4 | 0.070 | 0.087 | 0.088 | 0.071 | 0.115 | 0.109 | 0.119 | 0.106 | 0.094 |
| 5 | 0.076 | 0.085 | 0.092 | 0.087 | 0.097 | 0.104 | 0.110 | 0.106 | 0.097 |
| 6 | 0.070 | 0.097 | 0.096 | 0.092 | 0.099 | 0.094 | 0.110 | 0.112 | 0.104 |
| 7 | 0.073 | 0.097 | 0.109 | 0.107 | 0.102 | 0.098 | 0.097 | 0.101 | 0.100 |
| 8 | 0.067 | 0.0107 | 0.118 | 0.127 | 0.090 | 0.081 | 0.091 | 0.091 | 0.090 |
| 9 | 0.062 | 0.112 | 0.131 | 0.141 | 0.090 | 0.070 | 0.066 | 0.072 | 0.059 |
| High | 0.050 | 0.138 | 0.163 | 0.171 | 0.051 | 0.010 | 0.009 | 0.117 | 0.001 |

We examine sorted portfolios in this section. Table F.1, F.2, and F.3 report summary statistics for the 25 Fama-French size and book-to-market sorted portfolios, the ten beta-sorted portfolios, and the 90 low-turnover anomaly portfolios used in Novy-Marx and Velikov (2016) and Kozak, Nagel, and Santosh (2018).

We first examine the standard 25 size and book-to-market sorted portfolios that are available from Kenneth French's on-line data library. The 25 portfolios display the usual monotonic pattern in mean returns along the size and book-to-market dimension that we try to explain.

We use the aforementioned test statistics to capture the cross-sectional goodness-of-fit of a regression model. In addition, we also include the standard GRS test statistic. However, our othogonalization design does not guarantee that the GRS test statistic of the baseline model stays the same as the test statistic when we add an othogonalized factor to the model. The reason is that, while the othogonalized factor by construction has zero impact on the cross-section of expected returns, it may still affect the residual covariance matrix. Since the GRS statistic uses the residual covariance matrix to weight the regression intercepts, it changes as the estimate for the covariance matrix. We think the GRS statistic is not appropriate in our framework as its use of the residual covariance matrix to weight the regression intercepts is no longer optimal and may distort the comparison between candidate models. Indeed, for two models that generate the same regression intercepts, the GRS test will favor the model that explains a smaller fraction of variance in returns in time-series regressions.[32] To avoid this potential distortion from the GRS approach, we focus on the two metrics previously defined that do not rely on a model-based weighting matrix.[33]

We start by testing whether any of the 14 factors is individually significant in explaining the cross-section of expected returns. Panel A in Table F.4 presents the results. The market factor appears to be the best among the candidate factors. It reduces the mean scaled absolute intercept by 61%, much higher than what the other factors deliver.

[32] To see this, the GRS statistic equals $\alpha'\Sigma^{-1}\alpha$, where $\alpha$ is the cross-section of alphas and $\Sigma$ is the residual covariance matrix. For simplicity, let us take $\Sigma$ to be a diagonal matrix. Then the higher the values of the diagonal elements of $\Sigma$ are, the lower the GRS statistic is.

[33] It should be emphasized that it is not the GRS test statistic itself, but rather the ill-advised use of the test for relative model comparison, that may lead to biased comparison of candidate models. See Fama and French (1993) for a related argument.

## Table F.4: **Fama-French 25 Portfolios, Equally Weighted Scaled Intercepts**

Test results on 14 risk factors using Fama-French size and book-to-market sorted 25 portfolios. (See Table I for the definitions of risk factors.). The baseline model refers to the model that includes the pre-selected risk factors. We focus on the panel regression model described in Section 2.2. The two metrics (i.e., $SI_{ew}^{m}$ and $SI_{ew}^{med}$), which measure the difference in equally weighted scaled mean/median absolute regression intercepts, are defined in Section 3.2. GRS reports the Gibbons, Ross and Shanken (1989) test statistic.

Panel A: Baseline = No factor

| | | single test | | | single test | | |
|---|---|---|---|---|---|---|---|
| Factor | $SI_{ew}^{m}$ | $5th$-percentile | $p$-value | $SI_{ew}^{med}$ | 5th-percentile | $p$-value | GRS |
| ***mkt*** | **-0.607** | [-0.340] | (0.002) | **-0.672** | [-0.333] | (0.000) | 4.290 |
| ***smb*** | -0.209 | [-0.243] | (0.072) | -0.108 | [-0.257] | (0.215) | 4.402 |
| ***hml*** | 0.189 | [-0.100] | (0.999) | 0.230 | [-0.110] | (0.997) | 4.050 |
| ***mom*** | 0.224 | [-0.108] | (0.998) | 0.256 | [-0.120] | (0.998) | 4.302 |
| ***skew*** | -0.014 | [-0.040] | (0.195) | 0.007 | [-0.053] | (0.731) | 4.454 |
| ***psl*** | 0.043 | [-0.038] | (0.946) | 0.054 | [-0.044] | (0.952) | 4.286 |
| ***roe*** | 0.504 | [-0.150] | (1.000) | 0.470 | [-0.144] | (0.999) | 4.919 |
| ***ia*** | 0.607 | [-0.157] | (1.000) | 0.637 | [-0.164] | (1.000) | 4.553 |
| ***qmj*** | 0.820 | [-0.275] | (0.990) | 0.806 | [-0.273] | (0.983) | 5.594 |
| ***bab*** | 0.036 | [-0.042] | (0.952) | 0.030 | [-0.055] | (0.908) | **3.718** |
| ***gp*** | -0.042 | [-0.037] | (0.039) | 0.026 | [-0.049] | (0.892) | 4.096 |
| ***cma*** | 0.450 | [-0.143] | (1.000) | 0.464 | [-0.155] | (0.999) | 4.238 |
| ***rmw*** | 0.268 | [-0.126] | (0.991) | 0.273 | [-0.124] | (0.987) | 4.325 |
| ***civ*** | -0.281 | [-0.140] | (0.000) | -0.283 | [-0.141] | (0.002) | 4.132 |
| | | multiple test | | | multiple test | | |
| | *min* | **[-0.368]** | (**0.003**) | | **[-0.373]** | (**0.000**) | |

Panel B: Baseline = ***mkt***

| | | single test | | | single test | |
|---|---|---|---|---|---|---|
| Factor | $SI_{ew}^{m}$ | 5th-percentile | $p$-value | $SI_{ew}^{med}$ | 5th-percentile | $p$-value |
| ***mkt*** | | | | | | |
| ***smb*** | -0.068 | [-0.174] | (0.251) | -0.007 | [-0.211] | (0.481) |
| ***hml*** | -0.434 | [-0.260] | (0.000) | -0.397 | [-0.302] | (0.009) |
| ***mom*** | 0.218 | [-0.071] | (0.999) | 0.210 | [-0.113] | (0.985) |
| ***skew*** | -0.116 | [-0.085] | (0.025) | -0.134 | [-0.117] | (0.039) |
| ***psl*** | -0.038 | [-0.034] | (0.040) | -0.135 | [-0.055] | (0.004) |
| ***roe*** | 0.375 | [-0.106] | (1.000) | 0.366 | [-0.137] | (0.998) |
| ***ia*** | -0.318 | [-0.168] | (0.001) | -0.262 | [-0.206] | (0.012) |
| ***qmj*** | 0.560 | [-0.134] | (1.000) | 0.898 | [-0.173] | (1.000) |
| ***bab*** | -0.442 | [-0.154] | (0.000) | -0.447 | [-0.179] | (0.000) |
| ***gp*** | 0.202 | [-0.087] | (1.000) | 0.200 | [-0.128] | (0.988) |
| ***cma*** | **-0.476** | [-0.196] | (0.000) | **-0.500** | [-0.225] | (0.000) |
| ***rmw*** | 0.055 | [-0.056] | (0.991) | 0.132 | [-0.119] | (0.962) |
| ***civ*** | -0.219 | [-0.094] | (0.001) | -0.099 | [-0.128] | (0.088) |
| | | multiple test | | | multiple test | |
| | *min* | **[-0.289]** | (**0.001**) | | **[-0.342]** | (**0.000**) |

Panel C: Baseline = ***mkt*** + ***cma***

| | | single test | | | single test | |
|---|---|---|---|---|---|---|
| Factor | $SI_{ew}^{m}$ | $5th$-percentile | $p$-value | $SI_{ew}^{med}$ | 5th-percentile | $p$-value |
| ***mkt*** | | | | | | |
| ***smb*** | -0.232 | [-0.353] | (0.171) | **-0.295** | [-0.454] | (0.188) |
| ***hml*** | 0.001 | [-0.136] | (0.657) | 0.013 | [-0.230] | (0.615) |
| ***mom*** | 0.091 | [-0.067] | (0.981) | 0.115 | [-0.139] | (0.930) |
| ***skew*** | 0.005 | [-0.058] | (0.654) | 0.093 | [-0.134] | (0.896) |
| ***psl*** | -0.027 | [-0.028] | (0.054) | 0.222 | [-0.069] | (0.992) |
| ***roe*** | 0.911 | [-0.128] | (1.000) | 1.271 | [-0.228] | (1.000) |
| ***ia*** | 0.382 | [-0.106] | (1.000) | 0.631 | [-0.181] | (1.000) |
| ***qmj*** | 1.381 | [-0.153] | (1.000) | 1.857 | [-0.242] | (1.000) |
| ***bab*** | 0.101 | [-0.069] | (0.991) | 0.080 | [-0.153] | (0.880) |
| ***gp*** | **-0.260** | [-0.073] | (0.061) | -0.084 | [-0.104] | (0.081) |
| ***cma*** | | | | | | |
| ***rmw*** | 0.561 | [-0.119] | (1.000) | 0.644 | [-0.188] | (1.000) |
| ***civ*** | -0.160 | [-0.100] | (0.013) | -0.214 | [-0.211] | (0.049) |
| | | multiple test | | | multiple test | |
| | *min* | **[-0.356]** | (**0.148**) | | **[-0.464]** | (**0.253**) |

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 F.4, the bootstrapped $5^{th}$ percentile of $SI_{ew}^{m}$ for the market factor is -0.340. 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 $5^{th}$ percentile, there is a percentage reduction in the mean scaled intercept of 34%. The actual factor reduces the mean scaled intercept by more than the $5^{th}$ percentile, so we declare it significant. More precisely, by evaluating the 61% reduction against the empirical distribution of $SI_{ew}^{m}$ for the market factor alone, the single-test $p$-value for the market factor is 0.002.

To address the multiple testing problem, 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 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 be 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 F.4 shows the results on multiple testing. In particular, the bootstrapped $5^{th}$ percentile of $SI_{ew}^{m}$ for the minimum statistic is -36.8%. By evaluating the 61% reduction against the empirical distribution of the minimum statistic for multiple testing, the $p$-value is 0.003. Therefore, the multiple-test $p$-value is below the 5% cutoff. We therefore also declare the market factor significant from a multiple testing perspective. Across the two metrics we consider, the market factor is the dominating factor and is significant at 5% level, both from a single-test and a multiple-test perspective.

One interesting observation based on Table F.4 is that the best factor that is selected may not be the one with the lowest single test $p$-value. For instance, in Panel A of Table F.4 and for $SI_{ew}^{m}$, the market factor is the first factor that we select despite a lower single test $p$-value for *civ*. On the surface, this happens mechanically because the minimum test statistic picks the factor that has the lowest $SI_{ew}^{m}$ (i.e., highest percentage reduction in the mean scaled absolute intercept), not its $p$-value. As a result, the market factor, which has a lower $SI_{ew}^{m}$, is favored over *civ*.

On a deeper level, should we use a minimum test statistic that depends on the $p$-values instead of the levels of the $SI_{ew}^{m}$'s? We think not. The use of $SI_{ew}^{m}$ allows us to put weight on both the economic as well as the statistical significance. This is especially important for our sequential selection procedure that incrementally identifies the group of useful factors. We give a higher priority to a factor that has a large reduction in absolute intercept while passing a certain statistical hurdle than a factor that has a tiny reduction in absolute intercept but having a very small $p$-value.[34]

After the market factor is declared significant, we continue to identify the second risk factor. This time, *cma* has a multiple testing $p$-value of 0.001 under $SI_{ew}^{m}$ and less than 0.001 under $SI_{ew}^{med}$, and is therefore declared significant. Notice that the performance of *hml* is close to that of *cma*. This is not surprising given that *cma* and *hml* are highly correlated (correlation of 0.71).

After *cma* is identified and included in the baseline model, we continue to search for the third factor. This time, *gp* and *smb* are the best performing factors among the remaining factors under $SI_{ew}^{m}$ and $SI_{ew}^{med}$, respectively. Overall, across the two test statistics, *smb* seems to be a better performing factor compared to *gp* as it is close to *gp* under $SI_{ew}^{m}$ and a lot better than *gp* under $SI_{ew}^{med}$. Nonetheless, neither *smb* nor *gp* is significant under multiple testing. We therefore terminate the search and conclude with a two-factor model, i.e., $mkt + cma$, for portfolios sorted on size and book to market value.

Overall, our results using equally weighted scaled regression intercepts confirm the idea that *mkt* and *cma*, a factor that is closely related to *hml*, are helpful in explaining the cross-section of expected returns of Fama-French 25 portfolios. This may not be surprising as *hml* and Fama-French 25 portfolios use the same characteristics to sort the cross-section of stocks. What is interesting in our results is that *cma* survives after *mkt* is included. *smb* does not. This suggests that either *smb* is not a risk factor that helps explain the cross-section of expected returns for the Fama-French 25 portfolios or the Fama-French 25 portfolios have limited power to identify *smb* as a risk factor. As we shall see later, the latter explanation seems more plausible.

Our findings seem to be at odds with past studies that also use the Fama-French 25 portfolios to test the market factor. Most of these studies rely on the two-stage Fama-MacBeth regression and find that slope estimates from the second stage regressions are not statistically different from zero. Hence, the market factor is not priced. If one plots the estimated portfolio average returns against the actual average returns, one will see a flat line instead of a 45-degree line as one would expect to see if CAPM holds. In our framework, the market factor is highly significant. Indeed, based on Panel A of Table F.4, the market factor single-handedly reduces the scaled absolute regression intercept by about 60%.

[34] Notice that a different leverage scaling of a factor (i.e., long-short portfolio return) to alter its volatility will not change the test statistics or the $p$-values. This is because we run time-series regressions on the factors. Factor loadings adjust for different scalings. For example, when *mkt* is used as the factor, suppose we have a beta estimate of 1.0 for a certain asset. When $2 \times mkt$ is used, the beta estimate will drop to 0.5, offsetting the scaling of *mkt*. Meanwhile, neither the regression intercept nor its significance will be affected by the scaling.

Compared to cross-sectional methods such as the Fama-MacBeth approach, we provide an alternative framework to select risk factors. When the factor model is correctly specified (i.e., the beta pricing model is true for the factor model being tested), both methods seem to work well from a statistical standpoint although our model seems more powerful in detecting useful risk factors.[35] When there is potentially model misspecification, we believe that our approach is likely advantageous over Fama-MacBeth when there are a large cross-section of assets. First, given the large $N$, extreme observations are inevitable at each point in time for cross-sectional methods and may have undue impact on the inference. In contrast, our approach looks at regression intercepts that are estimated over a long period of time and is therefore less affected by extreme observations. In addition, we allow value weighting, which further alleviates the concern with outliers for small stocks.[36] Second, our method identifies reduction in regression intercepts offered by the augmented model relative to the baseline model. As a result, the additional factor in the augmented may be declared useful even if there is still a large regression intercept (in absolute value) left explained by the augmented model, capturing omitted factors and model misspecification. In contrast, cross-sectional methods, by having a common regression intercept at each point in time, are forcing the cross-section of expected returns to be completely driven by the factor model, and is therefore more likely to be affected by model misspecification.

While our results based on Fama-French 25 portfolios are interesting, we are reluctant to offer any deeper interpretation given the main drawback of the portfolio approach: tests based on characteristics-sorted portfolios are likely to be tilted towards factors that are constructed using the same characteristics.[37]

The GRS test statistic is problematic in our context from a variety of perspectives. For instance, with *mkt* as the only factor in the baseline model and by adding the orthogonalized *smb* to the baseline model, the GRS is 6.039 (not shown in table), much larger than 4.290 in Panel A of Table F.4, which is the GRS with *mkt* as the only factor. This means that by adding the orthogonalized *smb*, the GRS becomes much larger. By construction, the orthogonalized *smb* has no impact on the regression intercepts. The only way it can affect the GRS is through the error covariance matrix. Hence, the orthogonalized factor makes the GRS larger by reducing the error variance estimates. This insight also explains the discrepancy between $SI^m_{ew}$ and the GRS in Panel A of Table F.4: *mkt*, which implies a much smaller mean absolute intercept in the cross-section, has a larger GRS than *bab* as *mkt* absorbs a larger fraction of variance in returns in time-series regressions and thereby putting more weight on regression intercepts compared to *bab*.

[35] See the appendix for a comparison of test power between our approach and the cross-sectional approach.

[36] We discuss extensions of our approach that cope with time-varying risk exposures in our on-line appendix.

[37] See our on-line appendix for results that are based on the value-weighted test statistics and Fama-French 25 portfolios.

The weighting in the GRS does not seem appropriate for model comparison when none of the candidate models is expected to be the true model, i.e., the true underlying factor model that fully explains the cross-section of expected returns. Between two models that imply the same time-series regression intercepts, it favors the model that explains *a smaller fraction* of variance in returns. This does not make sense. We choose to focus on our proposed metrics that do not depend on the error covariance matrix estimate.

The way that the GRS test uses the residual covariance matrix to scale regression intercepts is likely to become even more problematic when we use individual stocks as test assets. Given a large cross-section and a limited time-series, the residual covariance matrix will be poorly measured. To make things worse, this covariance matrix needs to be inverted to obtain the weights for intercepts. As a result, the GRS test is likely to be very unstable and potentially distorted when applied to individual stocks.[38]

Our results about the GRS test resonate with a recent study by Fama and French (2015b). They find that the GRS test often implies unrealistically large short positions on certain assets, which does not make economic sense. To explain their findings, notice that the GRS test can be interpreted as the difference between the Sharpe ratio constructed using both the left-hand side assets and the right-hand side factors (call this Sharpe ratio $SR_1$) and the Sharpe ratio using only the right-hand side factors (call this Sharpe ratio $SR_2$). A rejection is found if $SR_1$ is significantly larger than $SR_2$. What Fama and French (2015b) find is that certain left-hand side assets need to take extreme short positions in order to achieve $SR_1$. By imposing short sale constraints, $SR_1$ is often much smaller, reducing the contribution of the left-hand side assets to the tangency portfolio formed using the right-hand side factors alone. This causes us to question the economic usefulness of the GRS test.

Our framework provides an economically meaningful approach to evaluate the incremental contribution of $SR_1$ over $SR_2$. In a panel regression model, the regression intercepts capture mispricing for the assets in the cross-section. An investor who is trying to exploit this mispricing will be long assets that have positive intercepts and short assets that have negative intercepts. By taking equally-weighted positions in the cross-section, the abnormal return for her portfolio (that is, returns with factor risks purged out) equals the equally weighted absolute intercepts plus a residual component that is the equally weighted average of the regression residuals. When we have a large cross-section — which will be the case when we use individual stocks as test assets — the residual component will be small. Therefore, the equally weighted absolute intercepts captures the abnormal return earned by an investor that tries to exploit the mispricing of the cross-section of assets relative to a factor model.

While the equally weighted absolute intercepts capture the abnormal returns of the equally weighted investment strategy, it is important to take the estimation un-

[38] See Gagliardini et al. (2014) for a similar argument.

certainty into account by using the standard errors to scale the regression intercepts. This motivates our test statistics (e.g., $SI_{ew}^{m}$) that are based on the scaled intercepts. As we mentioned before, our test statistics have substantially higher test power compared to tests that are based on the original intercepts.[39] Finally, an average investor in the economy will invest in proportion to the market capitalizations of assets. Hence, a value-weighted metric may better reflect the economic significance of asset mispricing in the cross-section. We explore this metric in the next section when we use individual stocks as test assets.

## F.2 Beta-sorted Portfolios

Are our results about the market factor specific to the Fama-French 25 portfolios we use? We construct the commonly used market beta-sorted portfolios to answer this question. In particular, we sort stocks into ten decile portfolios based on rolling beta estimates using information in the past 60 months. Table F.2 reports the summary statistics.

Applying our approach to the beta-sorted portfolios, we still first uncover the importance of the market factor, which singlehandedly reduce the scaled mean (median) absolute regression intercept by 55.3% (54.6%). After the market factor is identified, the next factor that stands out is betting-against beta (i.e., *bab*) — a factor that is the opposite to the beta strategy. We look into our results and find that after the market factor is included in the regression analysis, the regression intercepts are negative for the high-beta portfolio and positive for the low-beta portfolio. The returns of these portfolios are thus consistent with the prediction of *bab*. As a result, our second-stage regression picks up *bab*. Overall, our results highlight the roles of both market beta-related factors in driving the returns of beta-sorted portfolios.

Our results offer a different perspective in thinking about the returns of beta-sorted portfolios in relation to the market beta. Whereas cross-sectional tests usually find weak evidence on the importance of the market beta using beta-sorted portfolios, we demonstrate the importance of the market factor from a time-series (or panel regression) perspective. We are not the first to notice this difference. For example, Bollerslev, Li, and Todorov (2016) use a panel regression setup to establish the significance of jump-related market betas. Pukthuanthong, Roll, and Subrahmanyam (2019) propose a protocol to test candidate factors and advocate the use of the panel regression approach with individual stocks. Their empirical analysis identifies the market factor as the predominant risk factor. More recently, Hasler and Martineau (2019a, b) use panel regressions and find strong evidence consistent with CAPM. Giglio and Xiu (2019) focus on cross-sectional tests but use an approach that alleviates model mis-specification. They find universal evidence for the pricing of the market beta among a myriad of characteristics-based portfolios. Other recent papers that rejuvenate the market factor as a risk factor using alternative testing methods

[39] See Appendix B for results on our simulation study that measures test power.

include Ahn, Conrad, and Dittmar (2009), Cosemans et al. (2015), and Berk and Van Binsbergen (2016). Our results therefore add more evidence to this expanding literature that documents the importance of the market factor.

Table F.5: **Beta-sorted Portfolios, Equally Weighted Scaled Intercepts**

Test results on 14 risk factors using Beta-sorted decile portfolios. (See Table I for the definitions of risk factors.). The baseline model refers to the model that includes the pre-selected risk factors. We focus on the panel regression model described in Section 2.2. The two metrics (i.e., $SI^{m}_{ew}$ and $SI^{med}_{ew}$), which measure the difference in equally weighted scaled mean/median absolute regression intercepts, are defined in Section 3.2.

| 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*** | **-0.553** | [-0.301] | (0.000) | **-0.546** | [-0.299] | (0.000) | | | | | | |
| ***smb*** | -0.227 | [-0.252] | (0.072) | -0.209 | [-0.235] | (0.076) | -0.135 | [-0.301] | (0.253) | -0.130 | [-0.278] | (0.253) |
| ***hml*** | 0.146 | [-0.087] | (0.980) | 0.108 | [-0.071] | (0.974) | -0.318 | [-0.148] | (0.000) | -0.409 | [-0.171] | (0.000) |
| ***mom*** | 0.272 | [-0.133] | (0.995) | 0.275 | [-0.125] | (0.992) | 0.361 | [-0.113] | (1.000) | 0.301 | [-0.125] | (0.994) |
| ***skew*** | -0.018 | [-0.044] | (0.181) | -0.036 | [-0.051] | (0.082) | -0.072 | [-0.082] | (0.064) | -0.108 | [-0.097] | (0.033) |
| ***psl*** | 0.039 | [-0.040] | (0.952) | 0.029 | [-0.037] | (0.930) | -0.006 | [-0.036] | (0.297) | -0.026 | [-0.041] | (0.104) |
| ***roe*** | 0.612 | [-0.199] | (0.998) | 0.526 | [-0.174] | (0.998) | 0.898 | [-0.184] | (0.000) | 0.670 | [-0.183] | (1.000) |
| ***ia*** | 0.493 | [-0.148] | (0.996) | 0.442 | [-0.137] | (0.997) | -0.246 | [-0.088] | (0.000) | -0.367 | [-0.102] | (0.000) |
| ***qmj*** | 0.779 | [-0.266] | (0.976) | 0.712 | [-0.248] | (0.973) | 1.056 | [-0.225] | (1.000) | 0.790 | [-0.217] | (0.999) |
| ***bab*** | -0.058 | [-0.047] | (0.036) | -0.094 | [-0.060] | (0.018) | **-0.439** | [-0.165] | (0.000) | **-0.504** | [-0.182] | (0.000) |
| ***gp*** | -0.015 | [-0.030] | (0.124) | -0.042 | [-0.033] | (0.035) | 0.088 | [-0.048] | (0.995) | 0.034 | [-0.046] | (0.935) |
| ***cma*** | 0.334 | [-0.129] | (0.994) | 0.314 | [-0.127] | (0.993) | -0.412 | [-0.120] | (0.000) | -0.444 | [-0.133] | (0.000) |
| ***rmw*** | 0.302 | [-0.156] | (0.979) | 0.232 | [-0.125] | (0.975) | 0.358 | [-0.149] | (1.000) | 0.174 | [-0.116] | (0.988) |
| ***civ*** | -0.390 | [-0.173] | (0.002) | -0.359 | [-0.164] | (0.002) | -0.343 | [-0.184] | (0.003) | -0.420 | [-0.194] | (0.001) |
| | | multiple test | | | multiple test | | | multiple test | | | multiple test | |
| | *min* | **[-0.345]** | **(0.000)** | | **[-0.334]** | **(0.000)** | *min* | **[-0.322]** | **(0.002)** | | **[-0.312]** | **(0.000)** |

| Panel C: Baseline = ***mkt+bab*** | | | | | | | Panel D: Baseline = ***mkt*** + ***bab+civ*** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 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*** | -0.431 | [-0.546] | (0.122) | -0.392 | [-0.563] | (0.164) | -0.272 | [-0.658] | (0.386) | -0.310 | [-0.710] | (0.385) |
| ***hml*** | -0.170 | [-0.189] | (0.070) | -0.189 | [-0.229] | (0.075) | -0.147 | [-0.166] | (0.061) | -0.166 | [-0.221] | (0.079) |
| ***mom*** | 0.546 | [-0.350] | (0.971) | 0.493 | [-0.391] | (0.933) | 1.254 | [-0.366] | (0.999) | 1.191 | [-0.413] | (0.994) |
| ***skew*** | -0.017 | [-0.108] | (0.297) | -0.016 | [-0.136] | (0.310) | -0.026 | [-0.110] | (0.245) | -0.002 | [-0.157] | (0.428) |
| ***psl*** | 0.013 | [-0.074] | (0.751) | 0.017 | [-0.098] | (0.763) | 0.158 | [-0.091] | (0.977) | 0.170 | [-0.131] | (0.950) |
| ***roe*** | 1.389 | [-0.412] | (0.987) | 1.219 | [-0.439] | (0.964) | 3.528 | [-0.493] | (1.000) | 3.243 | [-0.576] | (0.997) |
| ***ia*** | -0.108 | [-0.103] | (0.044) | -0.211 | [-0.134] | (0.023) | -0.100 | [-0.100] | (0.049) | -0.291 | [-0.145] | (0.004) |
| ***qmj*** | 1.703 | [-0.421] | (0.991) | 1.460 | [-0.464] | (0.967) | 4.323 | [-0.543] | (1.000) | 3.881 | [-0.595] | (1.000) |
| ***bab*** | | | | | | | | | | | | |
| ***gp*** | 0.145 | [-0.096] | (0.986) | 0.054 | [-0.127] | (0.893) | 0.368 | [-0.119] | (0.996) | 0.317 | [-0.157] | (0.981) |
| ***cma*** | 0.385 | [-0.179] | (0.001) | -0.397 | [-0.223] | (0.009) | **-0.390** | [-0.176] | (0.002) | **-0.561** | [-0.244] | (0.002) |
| ***rmw*** | 0.450 | [-0.339] | (0.930) | 0.356 | [-0.377] | (0.880) | 1.158 | [-0.403] | (0.994) | 1.012 | [-0.448] | (0.977) |
| ***civ*** | **-0.619** | [-0.399] | (0.008) | **-0.629** | [-0.448] | (0.013) | | | | | | |
| | | multiple test | | | multiple test | | | multiple test | | | multiple test | |
| | *min* | **[-0.601]** | **(0.037)** | | **[-0.651]** | **(0.062)** | *min* | **-0.673** | **(0.394)** | | **-0.734** | **(0.246)** |

## F.3 Low-turnover Anomaly Portfolios

Finally, we consider 90 anomaly portfolios that are used in Novy-Marx and Velikov (2016) and Kozak, Nagel, and Santosh (2018). These portfolios have low turnover rates and thus more likely driven by risk-based explanations than mispricing. We report summary statistics on these portfolios in Table F.3. Note that these 90 portfolios, while overlapping with the 14 factors we study, do not span our factor space. As a result, one cannot use the mean-variance spanning test in Barillas and Shanken (2017) to evaluate the 14 factors. Our main goal for using these portfolios is to show how our results change in comparison with the previous two sets of portfolios.

Applying our approach to the 90 portfolios, we extract three factors: the market factor, common idiosyncratic risk (*civ*) and investment (*cma*). While *civ* is also significant based on our tests with beta-sorted portfolios, it is not under Fama-French 25 portfolios. In contrast, *cma* is declared significant under Fama-French 25 portfolios but disappears in tests with beta-sorted portfolios.

Overall, while there seems to be a core set of factors (i.e., market, *cma*, *bab*, *civ*) that underlie the three sets of anomaly portfolios we study, test results are sensitive to the construction of the portfolios. This is consistent with previous work such as Brennan, Chordia, and Subrahmanyam (1998). We therefore focus on tests with individual stocks in the subsequent sections.

Table F.6: **Low-Turnover Anomaly Portfolios, Equally Weighted Scaled Intercepts**

Test results on 14 risk factors using 90 low-turnover anomaly portfolios. (See Table I for the definitions of risk factors.). The baseline model refers to the model that includes the pre-selected risk factors. We focus on the panel regression model described in Section 2.2. The two metrics (i.e., $SI^{m}_{ew}$ and $SI^{med}_{ew}$), which measure the difference in equally weighted scaled mean/median absolute regression intercepts, are defined in Section 3.2.

| | 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*** | **-0.554** | [-0.319] | (0.001) | **-0.540** | [-0.317] | (0.001) | | | | | | |
| ***smb*** | -0.256 | [-0.248] | (0.041) | -0.266 | [-0.252] | (0.033) | -0.144 | [-0.271] | (0.232) | -0.154 | [-0.281] | (0.241) |
| ***hml*** | 0.218 | [-0.112] | (0.993) | 0.201 | [-0.121] | (0.982) | -0.236 | [-0.106] | (0.001) | -0.269 | [-0.105] | (0.001) |
| ***mom*** | 0.304 | [-0.147] | (0.997) | 0.298 | [-0.148] | (0.991) | 0.354 | [-0.118] | (1.000) | 0.312 | [-0.136] | (0.999) |
| ***skew*** | -0.005 | [-0.043] | (0.409) | -0.002 | [-0.045] | (0.468) | -0.042 | [-0.051] | (0.069) | -0.040 | [-0.058] | (0.106) |
| ***psl*** | 0.036 | [-0.034] | (0.942) | 0.033 | [-0.035] | (0.933) | -0.022 | [-0.031] | (0.092) | -0.047 | [-0.044] | (0.043) |
| ***roe*** | 0.673 | [-0.194] | (1.000) | 0.643 | [-0.197] | (0.999) | 0.861 | [-0.183] | (1.000) | 0.869 | [-0.207] | (1.000) |
| ***ia*** | 0.631 | [-0.161] | (1.000) | 0.593 | [-0.159] | (0.999) | -0.146 | [-0.054] | (0.001) | -0.166 | [-0.067] | (0.003) |
| ***qmj*** | 0.856 | [-0.284] | (0.986) | 0.825 | [-0.285] | (0.978) | 1.007 | [-0.206] | (1.000) | 0.986 | [-0.230] | (1.000) |
| ***bab*** | 0.072 | [-0.043] | (0.991) | 0.041 | [-0.044] | (0.961) | -0.231 | [-0.079] | (0.000) | -0.275 | [-0.086] | (0.000) |
| ***gp*** | -0.039 | [-0.036] | (0.040) | -0.046 | [-0.040] | (0.040) | 0.054 | [-0.036] | (0.986) | 0.025 | [-0.043] | (0.900) |
| ***cma*** | 0.439 | [-0.146] | (1.000) | 0.433 | [-0.151] | (0.996) | -0.300 | [-0.087] | (0.000) | -0.302 | [-0.094] | (0.000) |
| ***rmw*** | 0.330 | [-0.163] | (0.995) | 0.330 | [-0.165] | (0.990) | 0.330 | [-0.125] | (1.000) | 0.285 | [-0.153] | (0.994) |
| ***civ*** | -0.372 | [-0.174] | (0.000) | -0.387 | [-0.174] | (0.000) | **-0.322** | [-0.171] | (0.001) | **-0.404** | [-0.209] | (0.001) |
| | | multiple test | | | multiple test | | | multiple test | | | multiple test | |
| | *min* | **[-0.351]** | **(0.001)** | | **[-0.352]** | **(0.001)** | *min* | **[-0.288]** | **(0.028)** | | **[-0.319]** | **(0.007)** |
| | Panel C: Baseline = ***mkt+civ*** | | | | | | Panel D: Baseline = ***mkt + civ+cma*** | | | | | |
| | | 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*** | 0.006 | [-0.274] | (0.509) | -0.006 | [-0.347] | (0.478) | -0.075 | [-0.364] | (0.382) | -0.151 | [-0.435] | (0.302) |
| ***hml*** | -0.180 | [-0.078] | (0.000) | -0.174 | [-0.099] | (0.005) | 0.010 | [-0.045] | (0.731) | 0.016 | [-0.075] | (0.724) |
| ***mom*** | 0.360 | [-0.089] | (1.000) | 0.537 | [-0.127] | (1.000) | 0.500 | [-0.124] | (1.000) | 0.620 | [-0.171] | (1.000) |
| ***skew*** | -0.036 | [-0.037] | (0.051) | -0.055 | [-0.053] | (0.048) | -0.034 | [-0.034] | (0.047) | -0.112 | [-0.059] | (0.007) |
| ***psl*** | -0.001 | [-0.018] | (0.343) | -0.015 | [-0.034] | (0.149) | 0.014 | [-0.028] | (0.864) | 0.008 | [-0.049] | (0.698) |
| ***roe*** | 1.035 | [-0.156] | (1.000) | 1.223 | [-0.194] | (1.000) | 1.734 | [-0.160] | (1.000) | 1.923 | [-0.220] | (1.000) |
| ***ia*** | -0.195 | [-0.063] | (0.000) | -0.165 | [-0.079] | (0.001) | 0.287 | [-0.090] | (1.000) | 0.372 | [-0.133] | (0.999) |
| ***qmj*** | 1.242 | [-0.175] | (1.000) | 1.494 | [-0.229] | (1.000) | 2.125 | [-0.182] | (1.000) | 2.283 | [-0.232] | (1.000) |
| ***bab*** | -0.205 | [-0.064] | (0.000) | -0.264 | [-0.082] | (0.000) | -0.118 | [-0.054] | (0.003) | -0.203 | [-0.082] | (0.001) |
| ***gp*** | 0.040 | [-0.025] | (0.986) | 0.081 | [-0.052] | (0.987) | -0.078 | [-0.042] | (0.015) | -0.123 | [-0.054] | (0.001) |
| ***cma*** | **-0.296** | [-0.079] | (0.000) | **-0.269** | [-0.096] | (0.000) | | | | | | |
| ***rmw*** | 0.329 | [-0.104] | (1.000) | 0.456 | [-0.145] | (0.999) | 0.599 | [-0.111] | (1.000) | 0.633 | [-0.158] | (1.000) |
| ***civ*** | | | | | | | | | | | | |
| | | multiple test | | | multiple test | | | multiple test | | | multiple test | |
| | *min* | **[-0.282]** | **(0.040)** | | **[-0.348]** | **(0.142)** | *min* | **-0.364** | **(0.370)** | | **-0.438** | **(0.315)** |

# G Other issues

## G.1 Time-varying Factor Loadings

Our application focuses on panel regressions with fixed factor loadings. Our setting is therefore analogous to the environment of the GRS test where asset returns are projected onto factor proxies with constant factor loadings. It is also related to the two-pass cross-sectional regression method with time-invariant factor loadings, see, e.g., Shanken (1992), Jagannathan and Wang (1998), Shanken and Zhou (2007), and Kan, Robotti, and Shanken (2013).

While unconditional models approximate the asset pricing environment in a simple fashion, the true model might be conditional. Therefore, it might seem reasonable to always use a conditional model when possible. This is not true. Even when the true model is conditional, estimation errors for conditional betas may outweigh the gain of correctly specifying the true model, rendering the inference less efficient than an unconditional model specification. See Ghysels (1998).

Having discussed the pros and cons of conditional and unconditional model specifications, we explore two extensions of our panel regression tests that can accommodate time-varying factor loadings. The first extension is to explicitly model the conditioning variables as functions of financial and macroeconomic variables, as in Shanken (1990), Ferson and Harvey (1991), and Lettau and Ludvigson (2001). This effectively introduces new factors that interact the original factors with financial and macroeconomic variables. Our method follows by testing these new factors in addition to the original factors.

The second extension is to use the adapted Fama-MacBeth framework that we laid out in Section 2.3. We show how to modify the Fama-MacBeth framework so that the null hypothesis — the average of the time-series slope coefficients is zero — is exactly achieved in sample. We can then use this framework to incrementally select the list of true risk factors, similar to what we do with panel regressions. This framework allows us to take time-varying factor loadings into account as in the original Fama-MacBeth approach.

In this paper, we focus on unconditional models and leave these extensions to future research.

## G.2 Stepwise Model Selection

Our method falls in the realm of stepwise model selection, in particular forward selection for which we sequentially build up the true factor model. Unlike traditional F-tests or $R^2$ procedures, we pay particular attention to the multiple testing issue, making sure that the Wilkinson and Dallal (1981) critique does not apply to our method.

Having said this, we are aware of the issue that the $p$-value of our overall procedure, however defined, is likely to be a complex function of the $p$-values of the individual steps. In our simulation study, we also sidestep this issue by only considering the incremental selection of the second risk factor after the market factor is pre-determined. Notice that at each step, our method proposes a self-contained hypothesis testing framework that tests the incremental contribution of a group of candidate factors to a set of pre-determined factors. The $p$-value represents the multiple testing adjusted statistical significance of the best candidate variable. We leave a more detailed simulation study of our framework, in particular its overall performance in terms of selecting the true model or a model that has a significant overlap with the true model, to future research.

Traditional model selection methods such as the Akaike information criterion (AIC) or the Bayesian information criterion (BIC) are unlikely to work in our framework. We have a large panel of assets. The number of parameters to be estimated is proportional to the size of the cross-section. As a result, the asymptotic approximations that are required by AIC and BIC are unlikely to hold in our framework. Our bootstrap-based method provides a convenient way to make inference for a limited sample.

An alternative approach to forward selection is backward selection. That is, we start with an overall factor model and sequentially eliminate redundant factors. Our method does not apply to backward selection. To see why this is the case, imagine that we have 30 candidate variables. Based on our method, each time we single out one variable and measure how much it adds to the explanatory power of the other 29 variables. We do this 30 times. However, there is no baseline model across the 30 tests. Each model has a different null hypothesis and we do not have an overall null.

Besides the technical difficulty of backward selection, we think that forward selection makes more sense for our application. For the selection of risk factors, as a prior, we usually do not believe that there should exist hundreds of variables explaining a certain phenomenon. Forward selection is consistent with this prior.

## G.3 Spurious Factors

Spurious factors in factor models refer to factors that have weak covariance with asset returns. As shown in Kan and Zhang (1999) and Bryzgalova (2014), spurious factors make the usual inference methods problematic as the risk premia in factor models are weakly identified. Kan and Zhang (1999) and Bryzgalova (2014) propose diagnostic tools as well as shrinkage methods to detect and test spurious factors.

Our testing framework departs from the usual two-stage Fama-MacBeth framework or the associated generalized methods of moments (GMM) approach. In particular, we do not need to use the differences in factor loadings in the cross-section to identify factor risk premia, which is the source of the identification problem studied in Kan and Zhang (1999) and Bryzgalova (2014). Our method uses the reduction in absolute regression intercept (i.e., mispricing) as the test metric to gauge the success of a factor model. A spurious factor, by having a factor loading that is close to zero, naturally implies a small reduction in regression intercept and is likely to be identified as a false risk factor in our framework.

## G.4 Factor Model Uncertainty and Model Misspecification

To better link our method to the GRS test, one can think of the baseline model as the null hypothesis and the augmented model as the alternative hypothesis. The GRS test hypothesizes that the augmented model is the true underlying factor model and tests the deviations of the absolute regression intercepts from zero under this hypothesis. In this sense, the GRS test works under the alternative hypothesis. In contrast, our method works under the null hypothesis. We assume that the baseline model performs as well as the augmented model, that is, the additional factor in the augmented model has a zero contribution to explaining the cross-section of expected returns. We test whether the augmented model improves on the baseline model under this assumption.

Due to the above difference in the testing framework, our model is likely to be more powerful in identifying risk factors that belong to the true underlying factor model. For example, suppose the baseline model is simply a constant and the augmented model has the market factor as the candidate risk factor. For a given set of test assets and under GRS, we will reject the GRS null hypothesis (that is, CAPM is the true factor model) since there likely exist other factors that also drive asset returns but do not enter our test. As a result, we reject CAPM and conclude that the market factor cannot fully explain the returns of the test assets. This tells us little about whether the market factor is a useful risk factor or not. In contrast, in our framework, we are likely to identify the market factor as useful since the augmented model significantly improves on the baseline model (that is, a constant) in explaining the cross-section of expected returns.

In general, given the existence of hundreds of factors that are potential candidates for risk factors, there is a large amount of uncertainty around the true underlying factor model. As a result, any given factor model is likely to be misspecified. The use of the GRS test is limited since almost all models will be rejected in the end. Our test is less sensitive to model misspecifications and allows one to sequentially build the factor list. It does not try to make a statement about the underlying true factor model as in the GRS test. However, it tells us which factors are likely to be the members of the underlying true factor model.

# H FAQ

- *Isn't weighting by market cap just another way of creating size portfolios? (Section 2)*

  In our paper, value weighting allows us to take into account the differential impact of a factor across size groups. It could be the significance (e.g., $t$-statistic) of the factor across size groups. This is different from the return differential across size groups that are caused by the size effect.

- *GRS sometimes implies excessive short positions in certain assets. Does our approach suffer from the same problem?*

  No. Our method evaluates the average (either equal weighting or value weighting) level of mispricing. Importantly, it does not try to find the optimal portfolio choice that exploits mispricing. Hence, excessive short position is not a concern in our framework.

- *Does the difference in the GRS statistic between two models also measure the reduction in the intercepts?*

  The difference in the GRS statistics is a function of the intercepts and the covariance matrices. Suppose model A and B have the same intercepts but B implies a larger covariance matrix (assuming a diagonal matrix). Then model B is considered a better model than A by the GRS statistic. One issue with the use of the GRS statistic to measure relative model performance is that omitted factors may have a large impact on the covariance matrix, not just the intercepts.

- *In Table II, why is the single test 5th-percentile -34% for the market factor and the multiple test 5th-percentile trivially different at -38%?*

  This is actually one intuitive feature of our approach. In bootstrapped samples, extreme negative reductions are oftentimes caused by the orthogonalized market factor. In other words, the crowding out effect is not evenly distributed across factors. The reduction for the orthogonalized market factor (although on average zero by construction) has more variation, which leads to more extreme negative values compared to other factors.

# References

Adler, R., R. Feldman and M. Taqqu, 1998, A practical guide to heavy tails: Statistial techniques and applications, *Birkhäuser.*

Affleck-Graves, J. and B. McDonald, 1990, Nonnormalities and tests of asset pricing theories, *Journal of Finance 44, 889-908.*

Ahn, D., J. Conrad and R. Dittmar, 2009, Basis assets, *Review of Financial Studies 22, 5133-5174.*

Ang, A., J. Liu, and K. Schwarz, 2016. Using stocks or portfolios in tests of factor models. *Working Paper.*

Asness, C., A. Frazzini and L. H. Pedersen, 2019, Quality minus junk, *Review of Accounting Studies 24, 34–112.*

Avramov, D., and T. Chordia, 2006. Asset pricing models and financial market anomalies, *Review of Financial Studies 19, 1001-1040.*

Bai, J., and S. Shi. 2011. Estimating high dimensional covariance matrices and its applications. *Annuals of Economics and Finance 12, 199–215.*

Barillas, F., and J. A. Shanken, 2017. Which Alpha? *Review of Financial Studies, 1316–1338.*

Barillas, F., and J. A. Shanken. 2018. Comparing asset pricing models. *Journal of Finance 73, 715–754.*

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

Beran, R., 1988, Prepivoting test statistics: A bootstrap view of asymptotic refinements, *Journal of the American Statistical Association 83, 682-697.*

Berk, J. B., 1995, A critique of size-related anomalies, *Review of Financial Studies 8, 275-286.*

Berk, J. B., 2000, Sorting out sorts. *Journal of Finance 55, 407–427.*

Berk, J. B., and J. H. van Binsbergen. 2016. Assessing asset pricing models using revealed preferences. *Journal of Financial Economics 119, 1–23.*

Berk, J. B., and J. H. van Binsbergen. 2015. Measuring skill in the mutual fund industry. *Journal of Financial Economics 118, 1–20.*

Bollerslev, T., S. Z. Li, and V. Todorov. 2016. Roughing up beta: Continuous versus discontinuous betas and the cross section of expected stock returns. *Journal of Financial Economics 120, 464–490.*

Bryzgalova, S., 2016, Spurious factors in linear asset pricing models. *Working Paper, LSE.*

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

Cederburg, S., and M. S. O'Doherty, 2015, Asset-pricing anomalies at the firm level. *Journal of Econometrics 186: 113–128.*

Chen, H., S. Chen, Z. Chen, and F. Li. 2019. Empirical investigation of an equity pairs trading strategy. *Management Science 65, 370–389.*

Chen, L., R. Novy-Marx, and L. Zhang. 2010. An alternative three-factor model. *Working Paper.*

Chib, S., X. Zeng, and L. Zhao. On comparing asset pricing models. 2019. *Journal of Finance, Forthcoming.*

Chordia, T., A. Goyal, and J. Shanken, 2015, Cross-sectional asset pricing with individual stocks: Betas versus characteristics, *Working Paper.*

Cochrane, J. H., 2005, Asset Pricing. *Revised Edition.* Princeton University Press, Princeton, NJ.

Cosemans, M., R. Frehen, P. C. Schotman, and R. Bauer. 2015. Estimating security betas using prior information based on firm fundamentals. *Review of Financial Studies: hhv131.*

Croce, M. M., T. Marchuk, and C. Schlag. 2019. The leading premium. *Working Paper.*

Davidson, R., and J. G. MacKinnon, 2000, Bootstrap tests: How many bootstraps? *Econometric Reviews 19: 55-68.*

DeMiguel, V., L. Garlappi, and R. Uppal. 2009. Optimal versus naive diversification: How inefficient is the $1/N$ portfolio strategy? *Review of Financial Studies 22: 1915–1953.*

DeMiguel, V., L. Garlappi, F. J. Nogales, and R. Uppal. 2009. A generalized approach to portfolio optimization: Improving performance by constraining portfolio norms. *Management Science 55: 798-812.*

Ecker, F. 2013. Asset pricing tests using random portfolios, *Working Paper, Duke University.*

Efron, B. 1987, Better bootstrap confidence intervals, *Journal of the American Statistical Associations 82, 171-185.*

Efron, B. and R. J. Tibshirani, 1993, *An Introduction to the Bootstrap.* New York: Chapman & Hall.

Ehsani, S., and J. T. Linnainmaa, 2019, Factor momentum and the momentum factor. *Working Paper.*

Elton, E. J., M. J. Gruber, and C. R. Blake. The adequacy of investment choices offered by 401 (k) plans. *Journal of Public Economics 90: 1299–1314.*

Fama, E. F., 2015, Cross-section versus time-series tests of asset pricing models, *Working Paper.*

Fama, E. F. and J. D. MacBeth, 1973, Risk, return, and equilibrium: Empirical tests, *Journal of Political Economy 81, 607-636.*

Fama, E. F. and K. R. French, 1993, Common risk factors in the returns on stocks and bonds, *Journal of Financial Economics 33, 3-56.*

Fama, E. F., 1998, Market efficiency, long-term returns, and behavioral finance, *Journal of Financial Economics 42, 283–306.*

Fama, E. F. and K. R. French, 2008, Dissecting anomalies. *Journal of Finance 63, 1653–1678.*

Fama, E. F. and K. R. French, 2010, Luck versus skill in the cross-section of mutual fund returns, *Journal of Finance 65, 1915-1947.*

Fama, E. F. and K. R. French, 2015a, A five-factor asset pricing model, *Journal of Financial Economics 116, 1-22.*

Fama, E. F. and K. R. French, 2015b, Incremental variables and the investment opportunity set, *Journal of Financial Economics 117, 470-488.*

Fama, E. F. and K. R. French, 2018, Choosing Factors. *Journal of Financial Economics 128, 234–252.*

Ferson, W. E., and Y. Chen, 2015, How many good and bad fund managers are there, really? *Working Paper, USC.*

Ferson, W. E., and S. R. Foerster. 1994. Finite sample properties of the generalized methods of moments in tests of conditional asset pricing models. *Journal of Financial Economics 36, 29–55.*

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

Ferson, W. E., S. Sarkissian, and T. Simin. 1999. The alpha factor asset pricing model: A parable. *Journal of Financial Markets 2, 49–68.*

Feng, G., S. Giglio, and D. Xiu. 2019. Taming the factor zoo: A test of new factors. *Working Paper.*

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

Frazzini, A. and L. H. Pedersen, 2014, Betting against beta, *Journal of Financial Economics 111, 1-25.*

Gagliardini, P., E. Ossola, and O. Scaillet, 2014. Time-varying risk premium in large cross-sectional equity datasets, *Econometrica 84, 985–1046.*

Geweke, J., and G. Zhou. Measuring the pricing error of the arbitrage pricing theory. *Review of Financial Studies 9: 557–587.*

Ghysels, E., 1998, On stable factor structure in the pricing of risk: Do time-varying betas help or not? *Journal of Finance 53, 549–573.*

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

Giglio, S., and D. Xiu. 2019. Inference on risk premia in the presence of omitted factors. *Working Paper.*

Graham, J. R., and C. R. Harvey, 2001, The theory and practice of corporate finance: Evidence from the field, *Journal of Financial Economics 60, 187–243.*

Green, R. C. and B. Hollifield, 1992, When will mean-variance efficient portfolios be well diversified? *Journal of Finance 47: 1785–1809.*

Green, J., J. R. Hand and X. F. Zhang, 2017, The Characteristics that provide independent information about average U.S. monthly stock returns, *Review of Financial Studies 30, 4389–4436.*

Hall, P., 1988, Theoretical comparison of bootstrap confidence intervals (with Discussion), *Annals of Statistics 16, 927-985.*

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

Hansen, L. P., and R. Jagannathan, 1997, Assessing specification errors in stochastic discount factor models. *Journal of Finance 52: 557–590.*

Hansen, P. R., 2009, In-sample fit and out-of-sample fit: Their joint distribution and its implications for model selection. *Preliminary version, April, 23:2009.*

Harvey, C. R., 2017. Presidential Address: The scientific outlook in financial economics, *Journal of Finance 72, 1399–1440.*

Harvey, C. R. and Akhtar Siddique, 2000, Conditional skewness in asset pricing tests, *Journal of Finance, 55, 1263-1295.*

Harvey, C. R., Y. Liu and H. Zhu, 2016, ... and the cross-section of expected returns, *Review of Financial Studies 29, 5-68.*

Harvey, C. R. and Y. Liu, 2014, Multiple testing in economics, *Working Paper.* SSRN: http://ssrn.com/abstract=2358214

Harvey, C. R. and Y. Liu, 2017, Luck vs. skill and factor selection, *The Fama Portfolio, 250–260*, John Cochrane and Tobias J. Moskowitz, ed., 250-260, Chicago: University of Chicago Press.

Harvey, C. R. and Y. Liu, 2020, False (and missed) discoveries in financial economics. *Journal of Finance, forthcoming.*

Harvey, C. R. and Y. Liu, 2019, A census of the factor zoo. *Working Paper.* SSRN: http://ssrn.com/abstract=3341728

Harvey, C. R., and G. Zhou, 1990, Bayesian inference in asset pricing tests. *Journal of Financial Economics 26, 221–254.*

Hasler, M., and C. Martineau, 2019a. The dynamic CAPM. *Working Paper.*

Hasler, M., and C. Martineau, 2019b. Does the CAPM predict returns? *Working Paper.*

Herskovic, B., B. T. Kelly, H. N. Lustig, and S. Van Nieuwerburgh, 2016, The common factor in idiosyncratic volatility: Quantitative asset pricing implications. *Journal of Financial Economics 119, 249–283.*

Hoberg, G., and I. Welch. 2009, Optimized vs. sort-based portfolios. *Working Paper.*

Horowitz, J. L., J. J. Heckman, and E. E. Leamer, 2001, *Handbook of econometrics, Volume 5, 1–3843.*

Hou, K., C. Xue, and L. Zhang, 2015, Digesting anomalies: An investment approach, *Review of Financial Studies 28: 650–705.*

Hou, K., C. Xue, and L. Zhang, 2018. Replicating anomalies. *Review of Financial Studies, Forthcoming.*

Hinkley, D. V., 1989, Bootstrap significance tests, In *Proceedings of the 47th Session of the International Statistical Institute*, Paris, 29 August - 6 September 1989, 3, 65-74.

Jagannathan, R., and Z. Wang, 1998, An asymptotic theory for estimating beta-pricing models using cross-sectional regression. *Journal of Finance 53, 1285–1309.*

Jagannathan, R., and T. Ma. 2003. Risk reduction in large portfolios: Why imposing the wrong constraints helps. *Journal of Finance 58: 1651–1683.*

Jegadeesh, N., J. Noh, K. Pukthuanthong, R. Roll, and J. Wang. 2017, Empirical tests of asset pricing models with individual stocks: Resovling the errors-in-variables bias in risk premium estimation. *Working Paper.*

Jeong, J., and G. S. Maddala, 1993, A perspective on application of bootstrap methods in econometrics, In G.S. Maddala, C.R. Rao, and H.D. Vinod (eds), *Handbook of Statistics*, Vol. 11. Amsterdam: North Holland, 573-610.

Kan, R., and C. Zhang, 1999, Two-pass tests of asset pricing models with useless factors. *Journal of Finance 54, 203–235.*

Kan, R., and C. Robotti, 2008, Specification tests of asset pricing models using excess returns. *Journal of Empirical Finance 15, 816–838.*

Kan, R., and C. Robotti, 2009, Model comparison using the Hansen-Jagannathan distance. *Review of Financial Studies 22: 3449-3490.*

Kan, R., C. Robotti, and J. Shanken, 2013, Pricing model performance and the two-pass cross-sectional regression methodology. *Journal of Finance 68, 2617–2649.*

Kandel, S., and R. F. Stambaugh. 1995. Portfolio inefficiency and the cross-section of expected returns. *Journal of Finance 50: 157–184.*

Kelly, Bryan T., S. Pruitt, and Y. Su. 2019. Characteristics are covariances: A unified model of risk and return. *Journal of Financial Economics 134, 501–524.*

Kleibergen, F. 2009, Tests of risk premia in linear factor models. *Journal of Econometrics 149: 149–173.*

Kleibergen, F., and Z. Zhan, 2014. Mimicking portfolios of macroeconomic factors. *Working Paper.*

Kogan, L., and M. H. Tian. 2017. Firm characteristics and empirical factor models: A model-mining experiment. *Working Paper.*

Kosowski, R., A. Timmermann, R. Wermers and H. White, 2006, Can mutual fund "stars" really pick stocks? New evidence from a bootstrap analysis, *Journal of Finance 61, 2551-2595.*

Kozak, S., S. Nagel, and S. Santosh. 2017. Interpreting factor models. *Journal of Finance 73, 1183–1223.*

Ledoit, O., and M. Wolf. 2003. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance 10: 603–621.*

Ledoit, O., and M. Wolf. 2004. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis 88: 365–411.*

Ledoit, O., and M. Wolf, 2017, Nonlinear shrinkage of the covariance matrix for portfolio selection: Markowitz meets goldilocks. *Review of Financial Studies 30, 4349–4388.*

Lettau, M., and S. Ludvigson, 2001, Consumption, aggregate wealth, and expected stock returns. *Journal of Finance 56, 815–849.*

Lewellen, J., S. Nagel and J. Shanken, 2010, A skeptical appraisal of asset pricing tests, *Journal of Financial Economics 96, 175-194.*

Li, Q. and G. S. Maddala, 1996, Bootstrapping time series models, *Econometric Reviews 15, 115-195.*

Litzenberger, R. H., and K. Ramaswamy. 1979. The effects of personal taxes and dividends on capital asset prices: Theory and empirical evidence. *Journal of Financial Economics 7: 163–195.*

Lo, A. W., and A. C. MacKinlay, 1990, Data-snooping biases in tests of financial asset pricing models, *Review of Financial Studies 3, 431–467.*

Ludvigson, S. C., and S. Ng. 2009. Macro factors in bond risk premia. *Review of Financial Studies 22, 5027–5067.*

Ludvigson, S. C., 2011, Advances in consumption-based asset pricing: Empirical tests. *Handbook of the Economics of Finance*, G. M. Constantinides and R. M. Stulz, (eds.), Vol. 2, 799–906. Elsevier, Amsterdam.

MacKinlay, A. C., 1987, On multivariate tests of the CAPM, *Journal of Financial Economics 18, 341-371.*

MacKinlay, A. C., and M. P. Richardson, 1991, Using Generalized Method of Moments to test mean-variance efficiency, *Journal of Finance 46, 511–527.*

MacKinnon, J. G., 2006, Bootstrap methods in econometrics, *Economic Record 82, S2-18.*

McLean, R. D. and J. Pontiff, 2016, Does academic research destroy stock return predictability? *Journal of Finance 71, 5-32.*

Michaud, R. O. 1989. The Markowitz optimization enigma: Is 'Optimized' optimal? *Financial Analysts Journal, January-February, 31–42.*

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

Novy-Marx, R., and M. Velikov. A taxonomy of anomalies and their trading costs, *Review of Financial Studies 29, 104–147.*

Pástor, L. and R. F. Stambaugh, 2003, Liquidity risk and expected stock returns, *Journal of Political Economy 111(3).*

Paulsen, D., and J. Söhl, 2016. Noise fit, estimation error and a Sharpe information criterion. *Working Paper.*

Plyakha, Y., R. Uppal, and G. Vilkov, 2016, Equal or value weighting? Implications for asset-pricing tests, *Working Paper.*

Politis, D. and J. Romano, 1994, The Stationary Bootstrap, *Journal of the American Statistical Association 89, 1303-1313.*

Pukthuanthong, K., R. Roll and A. Subrahmanyam, 2019, A protocol for factor identification, *Review of Financial Studies 32, 1573–1607.*

Roll, R. 1977. A critique of the asset pricing theory's tests Part I: On past and potential testability of the theory. *Journal of Financial Economics 4: 129–176.*

Ross, S. A. 1976. The arbitrage theory of capital asset pricing. *Journal of Economic Theory 13: 341–360.*

Shanken, J., 1985, Multivariate tests of the zero-beta CAPM. *Journal of Financial Economics 14, 327–348.*

Shanken, J., 1987, Multivariate proxies and asset pricing relations: Living with the Roll critique. *Journal of Financial Economics 18: 91–110.*

Shanken, J., 1990, Intertemporal asset pricing: An empirical investigation, *Journal of Econometrics 45, 99–102.*

Shanken, J., 1992, On the estimation of beta-pricing model, *Review of Financial Studies 5, 1–33.*

Shanken, J., and G. Zhou, 2007, Estimating and testing beta pricing models: Alternative methods and their performance in simulations. *Journal of Financial Economics 84, 40–86.*

Sharpe, W. F., 1964, Capital asset prices: A theory of market equilibrium under conditions of risk, *Journal of Finance 19, 425–442.*

Stambaugh, R. F., and Y. Yuan. 2016. Mispricing factors. *Review of Financial Studies 30, 1270–1315.*

Sullivan, S., A. Timmermann, and H. White, 1999, Data-snooping, technical trading rule performance, and the bootstrap, *Journal of Finance 54, 1647-1691.*

Treynor, J. L., and F. Black, 1973, How to use security analysis to improve portfolio selection. *Journal of Business, 66-86.*

Uppal, R., and P. Zaffaroni, 2016, Portfolio choice with model misspecification: A foundation for alpha and beta portfoios. *Working Paper.*

Veall, M. R., 1992, Bootstrapping the process of model selection: An econometric example, *Journal of Applied Econometrics 7, 93-99.*

Veall, M. R., 1998, Applications of the bootstrap in econometrics and economic statistics, In D.E.A. Giles and A. Ullah (eds.), *Handbook of Applied Economic Statistics.* New York: Marcel Dekker, chapter 12.

White, H. Halbert, 2000, A reality check for data snooping, *Econometrica 68, 1097-1126.*

Wilkinson, L., and G. E. Dallal, 1981, Tests of significance in forward selection regression with an F-to-enter stopping rule, *Technometrics 23, 377-380.*

Xiaoji, L., B. Palazzo, and F. Yang. 2019. The risks of old capital age: Asset pricing implications of technology adoption. *Journal of Monetary Economics, Forthcoming.*

Young, A., 1986, Conditional data-based simulations: Some examples from geometrical statistics, *International Statistical Review 54, 1-13.*