Empirical Decisions and Replicating Anomalies

Jiaqi Guo[a] *and Peng Li*[b]

This version: August 2023

Abstract

There is an ongoing debate about the reproducibility of anomalies and *p*-hacking (data mining) of anomaly discoveries. This paper simulates and evaluates the impact of empirical decisions on anomaly replication and *p*-hacking. To better capture the true anomaly effect, we aggregate its return across 96 portfolio construction designs, avoiding dependence on any particular design. We develop a two-stage bootstrap approach to account for both sampling and empirical decision variations. Our simulations show that 70% of the published 173 anomalies can be replicated and the aggregate method in computing anomaly returns in the actual data is robust to both types of errors, suggesting that researchers should use the aggregate method to discover new anomalies, which alleviates the concerns of p-hacking through different empirical choices. Furthermore, we simulate anomaly discoveries through *p*-hacking activities and publication bias behavior. The findings indicate the existence of *p*-hacking attempts especially when the *t*-value threshold is 2 but the extent of *p*-hacking is not severe in anomaly studies.

Keywords: Empirical decisions, Anomaly replication, Bootstrap simulation, *p*-hacking, Data mining

JEL: C58, G10, G11, G12

[a] Jiaqi Guo, Birmingham Business School, University of Birmingham, Edgbaston, Birmingham, B15 2TT, United Kingdom. E-mail: j.guo.3@bham.ac.uk

[b] Peng Li, School of Management, University of Bath, Claverton Down, Bath, BA2 7AY, United Kingdom. E-mail: pl750@bath.ac.uk

1. Introduction

Over the past several decades, numerous anomalies or factors are documented to have the capacity to capture cross-sectional return pattern. For example, buying past winner stocks and selling short past loser stocks can generate significant returns (Jagadeesh and Titman, 1993). The 'factor zoo' has continuously expanded over time. Harvey and Liu (2019) summarize more than 400 factors, and Chen and Zimmerman (2022) review and replicate more than 200 existing anomalies. Given the proliferation of anomalies and replication crisis experienced in other research fields[1], there is growing attention on anomaly replication and potential *p*-hacking (data mining) in asset pricing studies in recent years (Harvey, Liu and Zhu, 2016; Mclean and Pontiff, 2016; Yan and Zheng, 2017; Hou, Xue and Zhang, 2020; Chen and Zimmerman, 2020; Jensen, Kelly and Pedersen, 2023).

The literature continues to debate three crucial questions. First, 'Can all published anomalies be replicated?'. Second, 'Is there *p*-hacking in anomaly discovery?'. Third, 'if the anomaly is genuine, how large is the true anomaly effect?'[2]. Hou, Xue and Zhang (2020) show that 65% of 452 anomalies cannot pass the 5% significance (t = 1.96) hurdle in their replication setting. Harvey, Liu and Zhu (2016) argue that many factors might be false due to publication bias or *p*-hacking. They suggest that multiple testing should be considered when evaluate the bunch of claimed factors, and the *t*-value should be raised to 3 to mitigate the false discovery rate. In contrast, Yan and Zheng (2017) indicate that at least the top-ranked fundamental signals exhibit significant returns, which are less likely to be driven by data mining. Chen and Zimmerman (2022) find that the majority of replicated *t*-values are close to the original studies, which confirms the credibility of discovered anomalies. Chen (2021) shows that the anomalies are unlikely to be the outcome of *p*-hacking.

[1] See Ioannidis (2005), Head et al. (2015), Camerer et al. (2016) and Chang, Gao and Li (2023).

[2] The true size of anomalies is paid little attention compared with *t*-statistics.

[insert a bar figure]

Anomalies' reproducibility can be affected by two forms of *p*-hacking. First, researchers can search among many variables and only report the significant one. Harvey and Liu (2021) and Chordia, Goyal, and Saretto (2020) have documented the evidence by simulating scenarios of searching different variables and find that around 50% anomalies are false positives. Second, researchers can experiment various empirical decisions (methods) and only present the (more) significant result. Menkveld et al. (2023) show that empirical results vary among different researchers in finance studies. In addition to data generating process (DGP) which leads to standard errors due to randomness in data samples, the variations across studies (non-standard error from evidence generating process(EGP) by different researchers) are important sources of uncertainties in empirical studies. Walter, Weber and Weiss (2023) reveal the size of non-standard error for anomaly returns using different methods of portfolio construction and show which empirical choices are the main drivers of the variations of anomaly returns. Hasler (2023) shows that method used in original paper produces higher anomaly return than alternative methods. Their study highlights the fact that there are non-negligible uncertainties when replicating anomaly returns. This highlights that empirical methods are an important source of statistical biases[3]. However, a crucial question arises: are there replication crisis and *p*-hacking in anomaly studies when viewed through the lens of empirical decisions?

This paper aims to answer this question by evaluating and quantifying the impact of empirical decisions on anomaly replication and *p*-hacking. To address this question, one should consider both "standard errors", i.e., sampling errors of individual study and "non-standard errors", i.e., specification errors arising from various empirical decision choices across studies. The observed anomaly long-short return should be the true effect plus two sources of errors:

[3] Other studies include Brodeur, Le, Sangnier and Zylberberg (2016) and Brodeur, Cook and Heyes (2020).

sampling and specification (study) errors (see Hedges and Vevea, 1998 and Thompson, Turner and Warn, 2001). For any individual anomaly, the observed anomaly return ($\hat{\theta}_k$) is the sum of the true anomaly return in study k (θ_k) and the sampling error (ε_k) from a sample drawn from the population:

$$\hat{\theta}_k = \theta_k + \varepsilon_k$$

Different studies have different replication rate……..

If all results based on different empirical decisions originate from a homogenous population, then different results can be attributed solely to the sampling error. However, if the true effects (θ_k) from different empirical specifications are heterogenous, it is difficult to conclude which empirical setting is the correct representation of the anomaly return[4]. Consequently, it would be unjust to ascertain the number of true anomalies and the replication rate based on any specific empirical choice. In light of this, it is imperative to account for the fact that the true effect of anomaly returns may conform a distribution. The true effect of anomaly under any method choice (θ_k) should be the sum of the overall (universe) true effect (μ) and the specification error between methods (ϑ_k):

$$\theta_k = \mu + \vartheta_k$$

Therefore, the observed anomaly return consists of the overall true effect and two sources of variance:

$$\hat{\theta}_k = \mu + \varepsilon_k + \ \vartheta_k$$

This suggests two implications. First, the overall effect μ should be a better measurement to quantify the true effect size of an anomaly. Second, the heterogeneity in empirical decisions leaves the room of data mining or *p*-hacking due to the publication bias in addition to the statistical bias from sampling error.

[4] Jensen, Kelly and Pedersen (2023) compare replication rates of their study and empirical settings from different papers in Figure 1. The replication rates range from 35% to 82.4%.

Our study uses 173 anomalies and 96 portfolio construction designs (i.e., empirical specifications) for each anomaly[5]. First, using permutation test, we show evidence that anomaly returns based on different empirical specifications are not from the same distribution for more than 50% anomalies. The finding suggests that the empirical findings of anomalies are likely to be affect by the empirical choices.

Second, we aggregate long-short returns across specifications by using average or median as the measurement of true effect size of anomaly returns which does not depend on any particular specification. We find that the reported returns from original methods are higher than returns based on the aggregate measure. Then we perform permutation to assess whether the true size of anomaly performance (aggregate return) follows the same distribution with the returns based on the original paper method. Our results reveal that the aggregate returns and original-method returns tend to come from the same distribution for most of anomalies albeit the original method return tends to be higher.

Third, more importantly, we propose a two-stage bootstrapping approach to take into account of both sampling and empirical specification variations. In the first stage of bootstrapping, we randomly draw monthly return with replacement for all anomalies simultaneously to preserve the cross-sectional correlation among anomalies (e.g., Harvey and Liu, 2021). Similarly, in the second stage of bootstrapping, we bootstrap specifications with replacement for all anomalies simultaneously. Overall, we find that about 70% of the published anomalies are true when we allow two sources of statistical biases. The 95% confidence interval of true anomalies is between 60% and 78%. We gain supporting evidence from the out-of-sample analysis using our two-stage bootstrapping approach. The performance of anomalies tends to persist both in pre and post out-of-samples by taking into account of the publication effect in addition to the sampling and specification errors.

[5] Some anomalies do not have all 96 specifications due to data issue.

Fourth, we simulate the publications bias and *p*-hacking behavior using our bootstrapping approach. We also perform a binomial test to assess the extent of *p*-hacking efforts. The intuition is that the probability of just below and above around *t*-value cut-off (e.g., 2) should be equal in the absence of *p*-hacking. The results suggest that there exists potential *p*-hacking efforts where researchers try to experiment different specifications when *t*-value is close to 2 but the *p*-hacking becomes unlikely when *t*-value is around 3.

At last, we simulate the *p*-hacking behavior induced by publication bias using our bootstrap approach. We find the FDR monotonically increases with the publication bias when *t*-value is 2 but it remains low (2%) when *t*-value cutoff raises to 3. In addition, we find the *t*-value of 2.6 is sufficient to limit the FDR below 5% when we are less confident about the published anomalies (60% anomalies are deemed to be true from our bootstrap approach) and there is an extreme publication bias of 90%.

Our study contributes to the existing literature on anomaly replication and *p*-hacking in several aspects. First, discovered anomalies from the original papers often use different methods. There is no standard or unified framework when investigating anomalies. Different researchers use different construction designs to test anomalies, which are determined by their individual decisions that vary across studies. Consequently, this leads to different replication outcomes that are not directly comparable. Indeed, the original papers of 173 anomalies included in our paper do not employ the same set of empirical choices. This raises a question what is the criteria for verifying a reliable anomaly. Recent studies suggest that the replication rates are influenced by empirical choices. For example, Hou, Xue and Zhang (2020) document a 65% failure rate of replicated anomalies (long-short deciles) with value-weighted returns (for controlling microcap stocks) and around 40% failure rate for equal-weighted returns. Jensen, Kelly and Pedersen (2022) apply terciles and capped-value-weighted returns (to control overweight for both extremely small and extremely large stocks) and find that about 56% of factors can be verified. This rate increases to 61.3% after removing insignificant anomalies from the original studies. However, Chen and Zimmerman (2022 and 2023) try to closely follow both the methods and time periods of the original studies and find that nearly all anomalies can be replicated. In our study, we address this issue by using a more proper true anomaly effect which is not dependent on any specification to justify the replication rate. Rather than using a specific empirical design to test the significance of anomaly returns, we suggest to use the aggregate measure (mean or median) across specifications, which is shown to be robust of sampling and specification errors.

Second, the observed anomalies suffer the problem of type I error (the probability that a positive findings is the result of incorrect rejection of true null hypothesis). The false discovery rate will be more apparently if *p*-hacking induced by publication bias exists. The behavior of searching significant findings will increase the probability of false discovery because true negative findings will be dropped. Harvey and Liu (2016) raise the concern and suggest a higher hurdle of *t*-value. The variations of long-short returns in different empirical specifications may leave more room for (unintended) *p*-hacking. *P*-hacking not only involves trying various variables or factors, but also trying different empirical choices. For example, researchers may experiment with different numbers of portfolios, NYSE break- points or the universe breakpoints and return weighting scheme to search for significant results. In this study, we simulate *p*-hacking behavior and try to quantify of the extent of the *p*-hacking efforts resulting from researchers experimenting different empirical procedures to report significant results.

Our paper is also different from related recent papers. Menkveld et al. (2023) introduce the term "non-standard errors" resulting from variations across researchers in generating evidence process. This adds a further error to the standard error in which the population parameter is estimated with an error in a random sample from a population. Walter, Weber and

Weiss (2023) examine 68 anomalies using different methodological choices in portfolio sorts and show that non-standard errors are larger than standard errors. Hasler (2023) investigates 92 predictors with various combinations of empirical decisions and shows that long-short returns based on the original paper's decision are higher than returns using other decisions. Instead, we focus on exploring and quantifying the consequences and implications of empirical decisions on replication crisis and *p*-hacking by taking account of both standard errors (sampling uncertainties) and non-standard errors (specification uncertainties). In addition, our study has a more comprehensive list of anomalies that includes 173 anomalies.

Harvey and Liu (2021) run a two-step bootstrapping approach by randomly selecting factors and shuffling months to evaluate how many observed anomalies are true. Their study simulates one aspect of *p*-hacking scenarios where researchers try many different factors and hide insignificant results. In contrast, our study simulate another aspect of *p*-hacking scenarios of trying different empirical specifications by applying the two-stage bootstrapping approach to account for variations for both sampling and specification errors. We aim to capture the true effect size of anomaly that is unbiased from specification errors and to examine whether published anomalies are likely to be the outcome of sampling and specification variations.

Our paper is organized as follows. We describe our data and how to construct different portfolio specifications in Section 2. In Section 3, we detail the methodologies including the procedures of two-stage bootstrapping. We present the results in Section 4. Section 5 concludes.

2. Data and portfolio constructions

We collect two sets of data from https://www.openassetpricing.com/data/ (Chen and Zimmerman, 2022). First, we download returns and *t*-values of long-short portfolios based on methods applied in original papers (the data also extends the OP sample period to December 2021). Second, we download all 204 available firm characteristics. We also collect data from

CRSP to complement 3 anomalies: firm size measured by market value, short-term reversal proxied by return in previous month and stock price. In total, there are 207 firm characteristics. There are 28 factors that have poor distribution to construct quintile or decile portfolios including 27 dummy variable type factors (values are 0 and 1). We, therefore, include 179 anomalies after removing the 28 anomalies. Further, we require at least 50 anomalies in each month and at least 240 months for any single factor. As a result, we have 173 anomalies to be included in our analysis. The start date of construction varies across anomalies due to the availability of firm characteristics, and the end date is December 2021. All firm characteristics are signed based on the characteristics-return relation. For example, there is a negative relation between firm size and stock return. Then for the long- minus-short return, we use the return of the bottom portfolio minus the return of the top portfolio. With the signed characteristics, we can always use top minus bottom as the long short spread. So an anomaly with a *t*-value greater than 2 (5% significance level) can be treated as a significant factor.

To test the long-short portfolio returns under different empirical designs, we employ six layers of empirical decisions. The six layers are identified by summarizing the methodologies of original papers. In the stock screening layer, there are 3 choices. One researcher can decide to use all stocks, stocks excluding NASDAQ listing or stocks excluding financial industry. In the liquidity layer, one can choose no filter or removal of stocks lower than 5 dollars. For portfolio breakpoints, the two options are all-universe and NYSE breakpoints. There are two choices when deciding the number of portfolios, deciles or quintiles. One can also choose stock returns with adjustment of delisting or not. Finally, either equal-weighted or value-weighted returns can be chosen. There are 3, 2, 2, 2, 2, 2 empirical choices across the six layers respectively, resulting in a total 3 × 2 × 2 × 2 × 2 × 2 = 96 specifications. We follow each of the 96 specifications to compute long-short returns in each month and take the time-series average as the anomaly return.

3. Empirical methods

3.1 Monthly return aggregation

We construct 96 specifications following the empirical decisions in Section 2 for each anomaly. Then we aggregate long-short return for each anomaly in each month as follows:

$$ret_{i,t}^{agg} = f_{agg}\left(ret_{i,1,t}^{s}, ret_{i,2,t}^{s}, \dots, ret_{i,96,t}^{s}\right)$$

$ret_{i,t}^{agg}$ is aggregate long-short return in of anomaly *i* in month *t*. $ret_{i,j,t}^{s}$ is the long-short return of anomaly *i* following specification *j* in month *t*. $f_{agg}(\cdot)$ either takes mean or median.

We finally compute the returns (ret_{i}^{agg}) of anomaly *i* by taking the average of time-series aggregate long-short return over *T* months:

$$ret_{i}^{agg} = \frac{1}{T}\sum_{t=1}^{T} ret_{i,t}^{agg}$$

3.2 Permutation tests

In this subsection, we perform two permutation tests to compare the difference of anomaly returns across specifications and the difference between the aggregate and original method returns. The rationale of using permutation tests is to assess whether the observed differences in data are statistically significant and not merely due to random sampling variation.

First, for each anomaly, we conduct a permutation test to examine whether anomaly returns from different portfolio constructions (i.e., specifications) have the same mean return. We first calculate the observed *F*-statistics using monthly returns across all specifications. Then we pool all monthly long-short returns from all specifications and randomly shuffle and reassign them among different specifications so that each specification contains randomly selected monthly long-short return that can be from any specification. The *F*-statistics is estimated using random samples. The intuition is that if returns from different specifications are from the same distribution and have the same mean, then randomly reassigning the returns among the different specifications should not affect the observed F-statistics. We repeat the simulation 1000 times and compute 1000 F-statistics. Then we can calculate the probability (p-value) of simulated F-statistics (F^i) equal or greater than the observed F-statistics (F^a) below.

$$p = P(F \geq F^a | H_0) = \frac{\sum_{i=1}^{n} I\left(F^i \geq F^a\right) + 1}{n + 1}$$

H_0 is that mean returns of all specifications are equal. I is the indicator function that equals to 1 if the condition is met, and n is the number of permutations. We apply the empirical adjustment by adding 1 for both numerator and denominator to avoid zero p-value. Lower p-value indicates the rejection of null hypothesis H_0 and suggest that the returns from different specifications are not from the same distribution and may have different means.

Next, we perform the second permutation test to assess if the true size of anomaly performance (measured by the aggregate return) has the same distribution with the return based on the original paper method. For each anomaly, we first compute the monthly average or median across all specifications and calculate the difference of sample means between the aggregated sample and original method sample, which represents the actual difference between the two sets of returns. Then, for each anomaly, we pool the aggregated monthly long-short returns with the monthly returns following original paper and apply random allocations of returns to the two groups and compute the mean difference in returns between the two groups. This procedure is repeated 1000 times and compute 1000 mean differences. The replicated returns and original-method returns should be from two different distributions if the p-value (proportion of differences between means which are at least as extreme as observed mean difference) is low. The p-value is estimated as following:

$$p = P\big((|D| \geq |D^a|)\big|H_0\big) = \frac{\sum_{i=1}^{n} I\left(\left|D^i\right| \geq |D^a|\right) + 1}{n + 1}$$

H_0 is that aggregate returns and original-method returns come from the same distribution. D^i is the difference between return based on aggregate method and return following original method from a random sample while D^a is the actual difference. We take absolute value to perform two-sided test.

3.3 Two-stage bootstrapping

To account for two sources of errors, we propose a two-stage bootstrapping approach. We first bootstrap the time-series data for all anomaly returns to allow sampling variations. To preserve the cross-sectional correlation[6] among different anomalies, we follow Harvey and Liu (2021) and Chen and Zimmerman (2023) to bootstrap the months with replacement for all anomalies at the same time in order to take into account the fact that monthly performance of anomalies might be correlated cross sectionally. In the second stage of bootstrapping, we allow variations of specifications by bootstrapping all specifications with replacement for all anomalies simultaneously. This ensures that the specifications are consistent across anomalies. In other words, each anomaly is subjected to the same set of empirical choices during the bootstrapping process. Moreover, this stage of bootstrapping helps simulate scenarios in which researchers try different specifications to obtain (more) significant results. Then we aggregate long-short returns across all specifications by taking either mean or median in each month, and calculate *t*-value of time-series aggregated returns. We simulate 100 times in each of the two bootstrapping steps, resulting in in a total of 10,000 (100*100) samples.

Using the two-stage bootstrapping, we conduct four main tests. First, we compute the cross-sectional mean return of 173 anomalies (using monthly returns based on either average or median aggregate) for each bootstrapped sample and then we obtain the average and 95% confidence interval of the 10,000 simulations. We also compute the 10 percentiles (from 10% to 90%) along with minimum (0% percentile) and maximum (100% percentile) of *t*-values (estimated by the aggregated long-short returns) across the 173 anomalies in each simulation.

[6] Similar technique is applied by Fama and French (2010) and Ben-David, Li, Rossi and Song (2022) when they capture the cross-sectional correlations of mutual fund returns in bootstrap.

Then we compute the 95% confidence interval of each *t*-value percentile based on the distribution of that percentile from 10,000 samples. With the 10,000 bootstrapped samples of each cross-sectional *t*-value percentile, we estimate the empirical cumulative distribution function (ECDF) as following

$$F_n(t) = \frac{1}{n}\sum_{i=1}^{n} I(t_i \leq t)$$

$I(\cdot)$ is an indicator function to assign 1 and else 0 if *t*-value of any percentile in a random sample is smaller than a fixed *t*-value. The ECDF enables us to present the distribution of bootstrapped *t*-value percentiles.

Second, to formally test if the *t*-value percentiles of anomalies are not the result of pure chance, we demean the monthly long-short returns and perform the two-step bootstrapping. The demeaned returns assume the anomaly returns to be zero, and this enables us to quantify how likely that the simulated cross-sectional percentile of *t*-value of the 173 anomalies is greater than the actual percentile of *t*-value if there is no predictability of the predictor. And the ECDF is estimated by the equation as shown below

$$F_n(t|H_0\colon no\ predictability) = \frac{1}{n}\sum_{i=1}^{n} I(t_i \leq t)$$

We show the bootstrapped distribution of cross-sectional percentiles of *t*-values under the null hypothesis of no predictability. More importantly, we can fit the ECDF to the actual percentiles of *t*-values, and then compute 1 - $F_n(t)$, the *p*-value, which represents the chance to observe at least as extreme of the actual *t*-value percentiles. In addition, we use the actual percentiles of *t*-value greater than 2 as the benchmark for large *t*-values to test how likely to observe the *t*-value of anomalies from simulated samples exceeding the benchmark, assuming zero anomaly return.

Third, we implement the bootstrap method to investigate out-of-sample performance of anomalies. Out-of-sample periods are defined as 3 and 5 years before the start of the original paper sample period and 3 and 5 years after the end of the original paper sample period.

Lastly, we evaluate *p*-hacking and the false discovery rate (FDR) of published anomalies. To detect the extent of *p*-hacking, following Andrews and Kasy (2019) and Brodeur, Cook and Heyes (2020), we perform the binomial test. The assumption is that the probability of just below and just above around *t*-value cut-off (usually 2) should be equal if there is no *p*-hacking or publication bias. Suppose there are N anomalies in a window around *t*-value cut-off and we observe that k_{obs} of N are significant anomalies. Under the null hypothesis that probability of significant findings is 0.5, the one-sided *p*-value of observing as or more extreme outcome is

$$P(k \geq k_{obs}) = \sum\nolimits_{k=k_{obs}}^{N} \binom{N}{k} p^k (1-p)^{N-k}$$

Our bootstrapping procedure with the average aggregate is the measurement without *p*-hacking, so we expect that the proportion of significant findings in any window should be 50%, and the *p*-value is greater than 5%. If the OP anomalies have no *p*-hacking, then similar results should be observed. In contrast, if there is *p*-hacking, we should observe *p*-values to be less than 5%.

Regarding the estimation of FDR, we first identify the percentage of true and false anomalies before the bootstrapping. We keep the monthly return intact for true anomalies and demean monthly return for false anomalies. In the second step of bootstrapping, instead of bootstrapping specifications for anomalies simultaneously and calculating aggregate return across specifications, we randomly select one specification to act as no publication bias. To simulate publication bias behavior, in each bootstrapped sample, we select one specification for each anomaly after we drop 70%, 80% and 90% of insignificant specifications to simulate the *p*-hacking (the higher the proportion we drop the more *p*-hacking effort). The intuition is that by dropping insignificant specifications, there would be a higher likelihood of selecting significant specifications for false anomalies. In turn, we artificially increase the probability of false anomalies becoming significant. For each of the 10,000 simulations, we compute FDR as the ratio of the number of significant anomalies that are actually false divided by the total number of significant findings.

4. Empirical results

4.1 Uncertainties of anomaly performance from different empirical designs

For each of the 173 anomalies, we use 96 different empirical specifications to compute the long-short returns and *t*-values. Figure 1 plots boxplot of returns and *t*-values across specifications for each anomaly. The plot is sorted by the mean return across all specifications of each anomaly. In Panel A, it is evident that that most of anomalies have long-short mean returns above zero. This is consistent with Walter, Weber and Weiss (2023) that the predictors tend to have positive premium in most of specifications. In addition, all anomalies exhibit large uncertainties in their return performance across different empirical designs. While some anomalies yield returns up to around 3% per month, they can also have return close to or even below zero. Panel B plots the boxplot of *t*-values for each anomaly in different specifications. Similarly, the mean of *t*-values across specifications indicates that most of anomalies can pass the 5% significance level ($t >= 2$) hurdle. However, the range of *t*-values shows that many anomalies can be insignificant in some certain specifications.

[Insert Figure 1 about here]

The sizeable range of returns and *t*-values have several implications. First, anomalies are sensitive to empirical designs, and their significance is dependent on empirical choices. Second, it provides researchers opportunities and incentives to try and search for the (more) significant one. Third, the uncertainty of returns and *t*-values does not necessarily mean that an anomaly is not reproducible. Some of anomalies are always significant across various empirical decisions although there are large variations across different portfolio construction methods. In Figure 2, we plot the distribution of averaged standard error across specifications and the non-standard error among different methods. For each anomaly, following Menkveld et al. (2022) and Walter, Weber and Weiss (2023), the non-standard error, which is used to measure uncertainty between methods, is defined as the 3rd quartile of anomaly returns minus the 1st quartile of anomaly returns from the 96 specifications. In total, there are 173 averaged standard errors and non-standard errors. Figure 2 shows that the anomalies exhibit similar standard error (sampling error), but the non-standard error (between-method error) displays more variability and extreme values. This finding confirms that the empirical design is another critical source to affect the observed anomaly returns.

[Insert Figure 2 about here]

Overall, the variations across different empirical choices imply that the anomaly replication rate may differ if different empirical decisions are made, and some of the published predictors could be the outcome of specific method picking rather than true anomalies.

4.2. Actual long-short returns across all specifications

For each anomaly, we follow 96 different specifications to construct time-series long-short returns. We aggregate long-short return by taking the average or median of all specifications in each month to generate the aggregate monthly returns. The anomaly returns are then calculated as the time-series average of aggregate returns. The aggregate anomaly return provides a more accurate representation of the true effect of anomaly performance, and anomalies are more likely to be true if they can be replicated using aggregate returns since they are less likely to be the result of specification searching. Table 1 summarizes the anomaly returns based on aggregate return. The mean return of 173 anomalies is 0.36% per month based on average-specification aggregation and 0.35% per month based on median-specification aggregation. Our aggregate return is similar with Hasler (2023) who reports an average of 0.34% per month with more empirical choices. This suggests that the additional empirical decisions do not have impact on our results. When using the method in the original papers, the mean return of all anomalies is 0.5% per month[7].

[Insert Table 1 about here]

To test whether the sample mean returns across specification for each anomaly are same, we conduct the permutation test and find that (at the 5% significance level, 55.5% of anomalies reject the null hypothesis that mean returns from different specifications are the same and from the same distribution. This finding indicates that more than half of anomalies have between-specification or between-study heterogeneity and those anomaly returns are likely to be affected by the empirical choices. Next, we run the permutation test to investigate whether the aggregate long-short return distribution is as same as that following original method. When using average aggregation, around 13.9% anomalies reject the null hypothesis of same distribution at the 5% significant level. The median aggregation exhibits the similar proportion of rejections. Taken together, our results indicate that although the reported returns are higher from original paper methods (published) than aggregate returns across specifications, the aggregate returns and original-method returns come from the same distribution for most of anomalies.

Following Chen, Lopez-Lira and Zimmermann (2023), we group anomalies into three categories, risk-based, mispricing and agnostic explanations. Specifications should have little effect on risk-based anomalies. However, mispricing induced returns should be influenced more by empirical choices especially weighting scheme since those returns can be more easily arbitraged away for large firms than small firms. As expected, we observe that more

[7] The long-short returns are constructed by following the empirical choices in the original papers, and the sample period is extended to the end of 2021.

mispricing-based anomalies than risk-based anomalies (56% vs. 48.5%) reject the hypothesis of same distribution of long-short returns across all specifications. More mispricing-based anomalies (15.4%) than risk-based anomalies (12.1%) have larger returns following original method. The results suggest that empirical decisions have relatively larger impact on returns of mispricing-based anomalies.

We follow Chen and Zimmermann (2022) to group them into accounting, trading and other groups. We find other anomalies which do not use CRSP and Compustat data has the lowest percentage of rejection. Specifically, 24% of the other anomalies reject the null hypothesis of having the same mean and being from the same distribution across all specifications and 4% (8%) reject the null hypothesis that average (median) aggregate anomaly return has same distribution as the original-method returns.

In Table 2, we present the results for 6 anomalies in detail. These anomalies are chosen because they are probably the most well-known and are used to construct factors in asset pricing models (Fama-French (FF) (1992) 3- and FF (2015) 5-factors; Carhart (1997) 4-factor and Hou, Xue and Zhang (2015) q-factor model). Specifically, the 6 anomalies are used to construct returns of size (size), value (BMdec), investment (AssetGrowth), profitability (GP and RoE) and momentum (Mom12m) factors. We also report the results for each of the 173 in Table A1. We first find that different specifications are likely to generate different long-short returns given by lower *p*-values of F-test from the permutation test for all of the 6 anomalies except for return on equity (RoE). Second, the premutation test of same distribution between the aggregate and original-method returns shows 5 out of the 6 *p*-values are larger than 0.05, suggesting the aggregate and original-method returns tend to share the same distribution. The only exception is asset growth anomaly, which has a lower aggregate return than the original method. Therefore, the anomaly variables used to construct factor returns are reproducible after taking different specifications into account, and there is no significant difference between the aggregate returns and returns using the original paper method.

[Insert Table 2 about here]

We further breakdown the replication performance using specification aggregation based on different empirical choices in Tables 3 and 4. The aggregation of monthly long-short return for each anomaly is based on specifications following a particular empirical decision rather than all specifications. Table 3 shows that the price filter (excluding stocks with price less than 5 or not), return weighting scheme and the number of portfolios have significant impact on replication results. For example, the average return of all anomalies for equal-weighted (EV) specifications regardless of other empirical decisions is 0.43%, while the counterparty of value-weighted (VW) is 0.29%. The averaged *t*-value across all EW specifications is 4.28 which is the double that of the VW specifications (averaged *t*-value is 2.2). In addition, 80% and 53% of anomalies have *t*-value greater than 2 for EW and VW respectively, and the proportions are 61% and 26% if we apply 3 as the *t*-value threshold. We also find that excluding NASDAQ stocks has relatively weak impact and other empirical choices (desilting return adjustment, NYSE breakpoints and exclusion of financial firms) have no significant influence on the replication rate. Finally, we find around 70% replication success under all decisions except for the value-weighted scheme. This echoes Hou, Xue and Zhang (2020) that many anomalies cannot be replicated when applying value-weighted returns. We also confirm these findings in Table 4 by the regression of *t*-values estimated from some certain specification on *t*-values of 173 anomalies following original empirical choices. The R^2 ranges from 0.63 to 0.82. These imply a strong correlation between averaged *t*-values across different specifications and *t*-values generated by the original method. In summary, most of anomalies seem to be replicated successfully even there are variations between specifications. Although the original choices result in larger anomaly returns than replicated aggregate returns, this should not be the evidence to deny the reproducibility of most of anomalies.

[Insert Table 3 about here]

[Insert Table 4 about here]

4.3 Reproducibility based on the two-stage bootstrapping

To verify the reproducibility of anomalies, , we perform the two-stage bootstrapping approach to account for both sampling and specification errors. By conducting 100 bootstraps of months and 100 bootstraps of specifications, a total of 10,000 bootstrapped samples of time-series long-short returns are generated for each anomaly. For each sample, we aggregate returns across all specifications using average or median in each month to compute time-series aggregate long-short returns. Therefore, there are 173 anomaly returns and *t*-values for each bootstrap sample, and we then compute average return and *t*-values of the 173 anomalies.

In Table 5 Panel A, we report the average statistics of 10,000 bootstrapped samples. The mean return of anomalies based on average aggregate is 0.36% per month and the average *t*-value is 3.45. More than 70% and 50% of anomalies are significant at *t*-value of 2 andl 3 respectively. The 95% confidence interval for return and *t*-values are 0.31% to 0.41% and from 2.93 to 3.93 respectively. Hence, taking both sampling and specification errors into account, the reproducibility of 173 anomalies is quite achievable. Similar results using median aggregate are shown in Panel B of Table 5. Since the aggregate measure better reflects the true effect of anomaly return, the reported cross-sectional mean of anomaly returns from the simulated distribution should be more reliable. Therefore, the cross-sectional mean return of anomaly is around 0.35% per month and 4.2% when annualized.

[Insert Table 5 about here]

Next, we investigate the replication rate by grouping anomalies based on their explanations and types. Table 6 reports the results. We find that the replication rates do not vary too much across the three categories of underlying explanations. Specifically, Panel A of Table 6 shows that the average replication rate of risk-based anomalies from the 10,000 simulations is 68.9%, while the replication rate of mispricing-based and agnostics anomalies are above 70%. In panel B of Table 6, the anomalies are categorized into accounting, trading and other anomalies. Anomalies constructed with accounting data are classified as accounting anomalies, and trading anomalies are those using pricing data. Anomalies using other data, like analyst data, are assigned into other group. Other group, which does not use CRSP and Compustat data, exhibits a lower replication rate (around 55%) compared with accounting and trading group (above 70%) when using the *t*-value threshold of 2. The replication rate of other group is below 40% when using the threshold of 3. This discrepancy is likely due to the lower quality data used in other group compared to CRSP and Compustat data. As a result, the replication rate of the other group is more susceptible to empirical choices, resulting in a lower average replication rate. Overall, our two-step bootstrapping approach shows that around 70% anomalies can be repeatedly replicated and are not dependent on certain specifications.

[Insert Table 6 about here]

At last, we evaluate whether anomalies, as measured based on aggregate return in the actual data, can be attributed to sampling and specification errors. To address this, we compare the cross-sectional distribution of anomaly returns in the actual data with that in the simulated data using the two-stage bootstrapping. We first show the percentiles of realized *t*-values of the average (median) aggregate anomalies in Table 7. Starting from the 30% percentile, the *t*-values are greater than 2. The bootstrapped 95% confidence interval for each percentile is also reported. Panel A of Figure 3 plots the empirical cumulative distribution of 3 selected percentiles at 30%, 50% and 90%. It is very common to observe those percentiles or even large *t*-values in the simulations.

[Insert Table 7 about here]

[Insert Figure 3 about here]

To formally test whether the actual cross-sectional percentiles of *t*-values (greater than 2) are just by pure chance due to sampling and specification variations, we use the two-stage bootstrapping with demeaned returns. The demeaned returns implies that we assume the anomalies returns are zero. This allows us to test the probability (*p*-value) of observing extreme outcomes for each percentile of *t*-values given that the anomaly return is zero. If the *p*-value is lower than 0.05, then we could conclude that it is less likely to observe the percentile just by pure chance. In other words, if we find very few simulations generate returns that are as extreme as those in the actual data, this would suggest that the sampling and specification variations are not the source of the significant anomaly return.

We compute the cross-sectional percentiles of *t*-values across 173 anomalies in each simulation and obtain 10,000 bootstrapped samples for each percentile. Panel B of Figure 3 plots the ECDF for 30%, 50% and 90% percentiles, indicating that it is impossible to observe the actual percentiles by sampling and specification variations if there is no return predictability. Table 7 shows that bootstrapped *p*-values are zero from 30% percentiles to 100% percentiles for both average and median aggregate procedures. The bootstrapped *p*-value is the percentage of simulations in which the corresponding simulated percentile of *t*-value is greater than the corresponding value in the actual data. Hence, none of the 10,000 random samples can generate the percentile of *t*-value as extreme as the actual percentile. Such evidence confirms that 70% of published anomalies cannot be explained by sampling and specification variations.

In addition, using the *t*-values from 30% percentiles and 100% percentiles, we test how likely to observe those *t*-statistics from the 173 anomalies in each bootstrapped sample. That is, we calculate the proportion of all anomalies (*p*-value) in each bootstrapped sample whose *t*-value exceeds the *t*-value of the percentile in the actual data. Then we calculate the average *p*-value from 10,000 samples. It can be seen from the last column of Table 7 that, for average aggregate method, the chance to observe *t*-value of at least 2.08, 2.70 and 3.03 are 2.27%, 0.37% and 0.12% respectively by allowing sampling and specification variations under the null hypothesis that anomalies have zero returns. And for even larger t *t*-values, the *p*-values are zero. We therefore reject the null. The results suggest that it is very difficult to observe the *t*-values purely by chance if the null hypothesis of zero anomaly return is true. This implies that researchers are almost impossible to produce those *t*-values by experimenting with different samples and empirical choices if they are false anomalies. For instance, more than half of anomalies (around 86 anomalies) have *t*-values greater than 3.03 in the actual data. One should experiment 833 (1/0.0012) samples to find only one sample (with specific months and specifications) with the *t*-value greater than 3.03. Suppose the 86 anomalies are discovered by 86 researchers, it means that each researcher should try samples and specifications 833 times and in total 71,638 (833*86) times collectively to produce those significant anomalies.

Overall, it is highly creditable that around 70% of published anomalies should be true and they are less likely to be identified by experimenting different samples and empirical decisions.

4.4 Out-of-sample performance

If anomalies are indeed true, they should not disappear in out-of-sample. In this section, we perform the two-stage bootstrapping analysis in the out-of-sample. As mentioned in methodology, we define a short period, 3 or 5 years before the start of sample and after the end of sample rather than using all available period in the pre-sample (before the start date of original paper) and post-sample (after the publication year). We use a shorter out-of-sample period for three main reasons. First if anomaly returns are generated purely by chance, the returns should fall dramatically in a short period before or after in-sample period. Second, since McLean and Pontiff (2016) document that anomaly decay is due to investor learning after the publication of anomalies, post-publication returns should be affected by the publication effect.

Third, the pre-sample should avoid the publication effect, but there are many other factors driving the anomaly returns in different sample periods. The anomaly generating environment is different between in-sample and far-away periods from the in-sample. Taken together, it seems to be more suitable to compare the in-sample and out-of-sample returns during a closer period to control for effects from various anomaly driving factors.

We perform the two-stage bootstrapping for in-sample and each of the two out-of-sample periods to obtain distributions of cross-sectional statistics, accounting for both sampling and specification errors. We then compare the distribution of cross-sectional mean return of anomalies between the in-sample and out-of-sample. We focus on the return distribution rather than *t*-value because the in-sample period is much longer than the out-of-sample period, and as a result, the standard error for the in-sample is much smaller than the out-of-sample. Panel A of Figure 4 plots the distribution of cross-sectional mean return of 173 anomalies from 10,000 simulated samples. The mean of cross-sectional mean returns for in-sample is 0.5% per month, and the in-sample distribution of mean returns is located on the right of post 3- and 5-year out-of-sample. The distributions are similar for post 3- and 5-year out-of-samples ,which center around 0.34% and 0.31% per month, respectively. Thus, the mean return in the post out-of-sample is around 60% of the in-sample mean.

[Insert Figure 4 about here]

Panel B of Figure 4 shows that the distribution of pre out-of-sample is closer to the in-sample distribution compared to the post out-of-sample, and the mean of pre out-of-sample distribution is around 0.36% per month which is 70% of the in-sample mean. The higher of mean returns in the pre-sample than the post-sample may be due to investor learning after the publication of anomalies learning (McLean and Pontiff, 2016).

To alleviate the effect of investor learning after the publication of anomalies, we separate anomalies into two groups that are published before and after 1993. In 1993, the SEC started to implement the Electronic Data Gathering, Analysis and Retrieval (EDGAR), which facilitates information dissemination and eases the investor learning (see Gibbons, Iliev and Kalodimos 2021; Goldstein, Yang and Zuo, 2023). Therefore, the decay due to the publication effect should be larger for anomalies published after 1993 during the 3- to 5-year periods after publication. As shown in Panel A of Figure 5, which plots the distribution for anomalies published after 1993, anomalies during the in-sample outperforms the out-of-sample but the out-of-sample still accounts for 60% of the mean returns of the in-sample. Panel B provides the distribution for anomalies published before 1993 when investor learning is not as easy as after 1993. The distribution of out-of-sample more overlaps with the distribution of in-sample (the mean returns are 0.6% and 0.4% per month for in-sample and out-of-sample respectively). As the returns from our two-step bootstrapping are less likely the results of data-mining, the out-of-sample performance seems to persist. After accounting for investor learning, other factors seem to contribute at most 30% (1-0.4%/0.6%) of anomaly returns.

[Insert Figure 5 about here]

4.5 *P*-hacking and false discovery rate

In this section, we investigate the extent of *p*-hacking among published anomalies, using the two-stage bootstrapping approach. We first calculate *t*-values for each of the 173 anomalies in each simulation run based on the average aggregate long-short return and then compute the average of *t*-value for each anomaly across 10,000 samples. This serves as the *t*-value distribution of 173 anomalies without *p*-hacking. To simulate the effort of *p*-hacking, we modify the second stage of bootstrapping by randomly selecting a fraction of anomalies only using specifications yielding higher *t*-values rather than all specifications. By doing so, we are mimicking potential *p*-hacking behavior, where researchers might selectively choose specific specifications that yield higher *t*-values to highlight significant results, increasing the probability of publication.

Panel A of Figure 6 plots the *t* distribution of 173 anomalies based on the OP method in which less than 17% anomalies are not significant and *t*-values from 2 to 2.58 are observed considerably more frequent than *t*-values just below 2. In contrast, Panel B, which plots the *t* distribution without *p*-hacking, shows that the proportion of insignificant anomalies increases substantially to 27% compared with the OP method in Panel A (17%), and there is a slightly higher proportion of anomalies whose *t*-values are from 2 to 2.58 than *t*-values that are marginally below 2. The findings in Panels A and B indicate a potential *p*-hacking effort where *t*-values around 2 have an abnormally larger probability for the OP specification. This is confirmed by our simulation of *p*-hacking behavior. We create two scenarios of *p*-hacking in Panels C and D, in which we randomly choose 10% (20%) anomalies and use 70% (60%) percentile return across specifications rather than all specifications for those anomalies. As such, we artificially increase the probability of observing larger returns and therefore higher *t*-values. We observe consistent evidence in Panels C and D, suggesting potential *p*-hacking efforts from researchers by experimenting various specifications.

[Insert Figure 6 about here]

Next, we perform the binominal test to assess whether the observed *t*-values of OP anomalies are around a threshold with equal probability. The results are reported in Table 8. We define the *t*-value windows around 2 using different distances. Under the assumption of no *p*-hacking, the probability to obtain *t*-values below and above 2 should be equal within the window around 2. We present the one-sided *p*-value to examine whether there is a tendency that the probability is higher to observe *t*-value greater than 2. We find that for simulated anomalies which are free of *p*-hacking, the percent of significant anomalies is around 50% for all windows and the *p*-values to observe *t*-value greater than 2 are all above 5%. Therefore we cannot reject the null hypothesis that the probability is 0.5, indicating no evidence of *p*-hacking for simulated anomalies through two-stage bootstrapping. However, for the OP anomalies, many windows exhibit around 70% significant anomalies. Further, many *p*-values are below 10% or 5% and therefore we can reject the null hypothesis, suggesting the presence of *p*-hacking for the *t*-value threshold of 2. When we adjust the *t*-value threshold to 3, *p*-values from all windows are above 30% for the OP anomalies. This suggests that there may be no motivation to engage in *p*-hacking when the *t*-value is large.

[Insert Table 8 about here]

Given that OP anomalies have potential *p*-hacking, we run the two-stage bootstrapping to estimate the false discovery rate (FDR) of the published anomalies. To compute FDR, we first need to identify how many anomalies are true. Our bootstrapping results suggest that 70% of anomalies should be true. The conservative estimate (lower bound of confidence interval) suggests that at least 60% anomalies are true. Next, we demean the monthly long-short returns of false anomalies whose *t*-values are below 30% percentiles of *t*-values based on the actual aggregate returns. In the two-stage of bootstrapping, we do not apply any aggregate. Instead, we pick up one specification randomly from all bootstrapped specifications for each anomaly. This is the sample without *p*-hacking as we do not intend to choose the significant one. As detailed in methodology section, we simulate different degrees of *p*-hacking behavior by excluding a fractional of insignificant specifications. For example, 70% publications bias means that 70% of insignificant specifications will be removed, and then we select one specification for each anomaly from the remaining specifications in each bootstrap sample. By doing this, we are able to estimate how many significant findings that are actually true and false given *p*-hacking effort.

Considering that 70% anomalies are true and a *t*-value threshold is 2, Panel A of Table 9 shows that the FDR is less than 3% when there is no any *p*-hacking effort on samples and specifications. However, the FDR exceeds 5% if *p*-hacking is moderate or high (publication bias is more than 70%). When *t*-value threshold increases to 3, the FDR can be restricted under

2%. Panel A of Figure 7 illustrates that *t*-value of 2.3 is sufficient to limit the FDR under 5% in the presence of high publication bias of 90% if we assume 70% of anomalies are true based on our two-stage bootstrapping approach. Panel B of Table 9 and Panel B of Figure 7 use the lower bound of 95% confidence interval form Table 5 where 60% of anomalies are deemed to be true. When the *t*-value cutoff is 2, the FDR is 4% without publication bias, indicating that the FDR is likely to be higher than 5% even the publication bias is moderate. Panel B of Figure 7 suggests that the *t*-value should raise to 2.6 to maintain the FDR below 5% if publication bias is high.

[Insert Table 9 about here]

[Insert Figure 7 about here]

5. Conclusions

In this study, we examine a comprehensive list of 173 published anomalies with 96 different specifications based on a range of possible empirical decisions. We compute anomaly returns and *t*-values based on the aggregated monthly long-short returns across various specifications by taking the average or median. This is a more proper measurement of the true anomaly effect which does not depend on any particular specification.

We propose a two-stage bootstrap approach, which allows for variations from both sampling and specification (i.e., standard and non-standard errors). We find that 70% of anomalies are not the outcome of the two sources of those variations. It is reasonable to infer that 70% of the published anomalies are true. Our simulations also find persistent out-of-sample evidence after excluding other effects such as investors learning due to the publication effect. The findings suggest that the replication crisis in anomalies/factors studies might not be severe, and a substantial number of significant results are creditable. We also find supportive evidence from our *p*-hacking and false discovery tests by using the two-stage bootstrap approach. Although the original papers do produce more significant results compared with the true effects from our methodology, our simulations reveals that only a small fraction of anomalies with some extent of *p*-hacking can match the OP results. Additionally, a *t*-value of 2.6 is sufficient to limit the FDR below 5% in the scenario that we are less confident about the published anomalies and there is an extreme publication bias.

References

Andrews, I., and M. Kasy. 2019. Identification of and correction for publication bias. *American Economic Review*, 109, 2766–2794.

Ben-David, I., J. Li, A. Rossi and Y. Song. 2022. What Do Mutual Fund Investors Really Care About?. *Review of Financial Studies*, 35, 1723-1774.

Brodeur, A., N. Cook and A. Heyes. 2020. Methods matter: p-hacking and publication bias in causal analysis in economics. *American Economic Review*, 110, 3634-3660.

Brodeur, A., M. Le, M. Sangnier and Y. Zylberberg. 2016. Star wars: the empirics strike back. *American Economic Journal: Applied Economics*, 8, 1-32.

Camerer, C. F., and others. 2016. Evaluating replicability of laboratory experiments in economics. *Science*, 351, 1433–1436.

Carhart, M. M. 1997. On persistence in mutual fund performance. *The Journal of finance*, *52*(1), 57-82.

Chang, X., H. Gao and W. Li. 2023. Discontinuous distribution of test statistics around significance thresholds in empirical accounting studies. Working Paper, SSRN.

Chen, A. Y.. 2021. The limits of p-hacking: some thought experiments. *Journal of Finance*, 76, 2447-2480.

Chen, A. Y., A. Lopez-Lira and T. Zimmermann. 2023. Peer-reviewed theory does not help predict the cross-section of stock returns. Working Paper, arXiv.

Chen, A. Y., and T. Zimmermann. 2020. Publication bias and the cross-section of stock returns. *Review of Asset Pricing Studies*, 10, 249-289.

Chen, A. Y., and T. Zimmermann. 2022. Open source cross-sectional asset pricing. *Critical finance review*, 11, 207-264.

Chen, A. Y., and T. Zimmermann. 2023. Publication Bias in asset pricing research. *Oxford Research Encyclopedia of Economics and Finance*, forthcoming.

Chordia, T., A. Goyal and A. Saretto. 2020. Anomalies and false rejections. *Review of Financial Studies*, 33, 2134-2179.

Fama, E. F., & French, K. R. 1992. The cross-section of expected stock returns. *the Journal of Finance*, *47*(2), 427-465.

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., & French, K. R. 2015. A five-factor asset pricing model. *Journal of financial economics*, *116*(1), 1-22.

Gibbons, B., P. Iliev and J. Kalodimos. 2021. Analyst information acquisition via EDGAR. *Management Science*, 67, 769-793.

Goldstein, I., S. Yang and L. Zuo. 2023. The real effects of modern information technologies: Evidence from the EDGAR implementation. *Journal of Accounting Research*, forthcoming.

Harvey, C., and Y. Liu. 2019. A census of the factor zoo. Working Paper, SSRN.

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

Harvey, C., and Y. Liu. 2021. Uncovering the iceberg from its tip: A model of publication bias and p-hacking. Working Paper, SSRN.

Hasler, M.. 2022. Looking under the hood of data-mining. Working Paper, SSRN.

Head M. L., L. Holman, R. Lanfear, A. T. Kahn and M. D. Jennions. 2015. The extent and consequences of p-hacking in science. *PLoS Biology*, 13, 1-15.

Hedges, L. V., and J. L. Vevea. 1998. Fixed- and random-effects models in meta-analysis. *Psychological Methods*, 3, 486–504.

Hou, K., Xue, C., & Zhang, L. 2015. Digesting anomalies: An investment approach. *The Review of Financial Studies*, *28*(3), 650-705.

Hou, K., C. Xue and L. Zhang. 2020. Replicating anomalies. *Review of Financial Studies*, 33, 2019-2133.

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

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

Jensen, T. I., B. Kelly and L. H. Pedersen. 2023. Is there a replication crisis in finance?. *Journal of Finance*, online publication, https://doi.org/10.1111/jofi.13249.

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

Menkveld, A. J., and others. 2023. Non-standard errors. *Journal of Finance*, forthcoming.

Thompson, S. G., and R. M. Turner and D. E. Warn. 2001. Multilevel models for meta-analysis, and their application to absolute risk differences. *Statistical methods in medical research*, 10, 375–392.

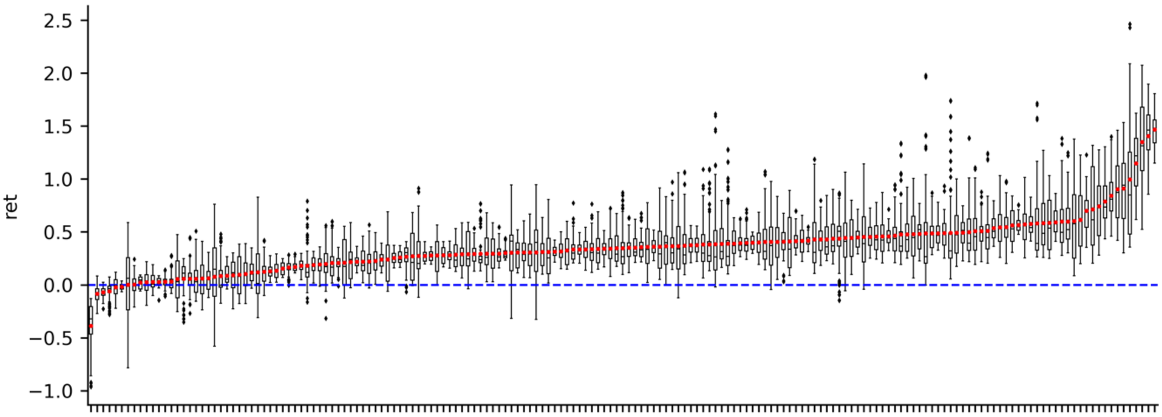
Walter, D., R. Weber and P. Weiss. 2023. Non-standard errors in portfolio sorts. Working Paper, SSRN.

Yan, X., and L. Zheng. 2017. Fundamental analysis and the cross-section of stock returns: A data-mining approach. *Review of Financial Studies*, 30, 1382-1423.

Figure 1. Returns and *t*-values of anomalies

The figure plots returns and *t*-values of anomaly for each specification. Panel A plots boxplot of returns for each anomaly across specifications. Panel B plots boxplot of *t*-values for each anomaly across specifications. Returns are time-series average of monthly long-short returns. There are 173 anomalies and 96 specifications. The red dot indicates the mean of all possible specifications.

Panel A: returns of across specifications



Panel B: *t*-values of across specifications

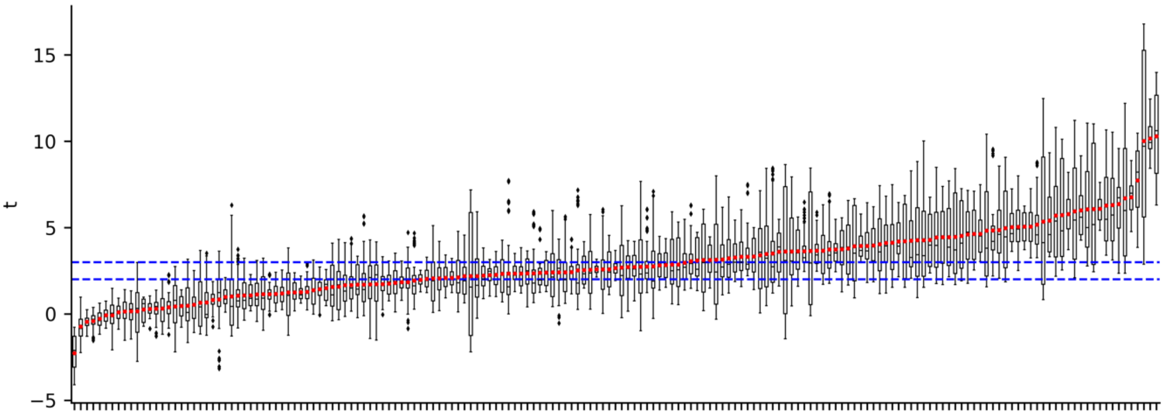


Figure 2. Distribution of standard and non-standard errors
The figure plots distribution of standard error and non-standard error. Long-short returns and standard error are computed in each of 96 specifications for each of 173 anomalies. Standard error of an anomaly is the average of standard errors across specifications. Non-standard error is the difference between 75% percentile of standard error and 25% percentile of standard error.

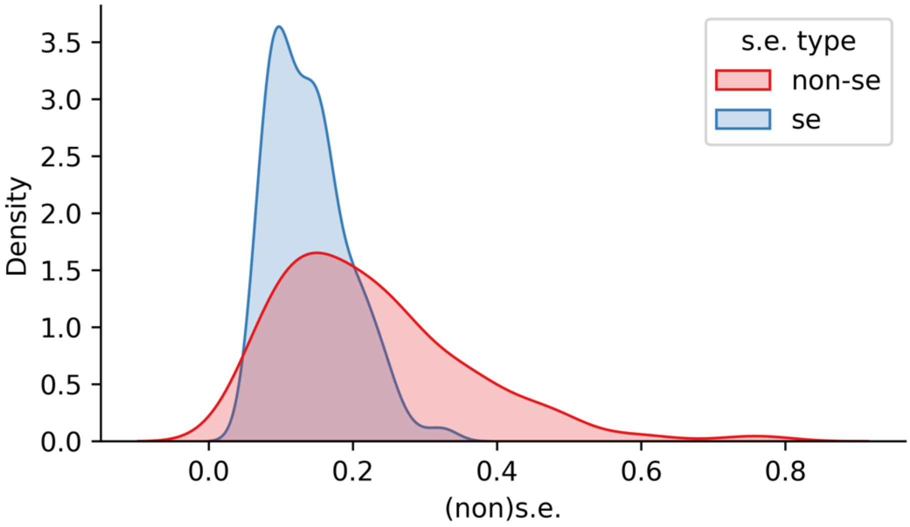
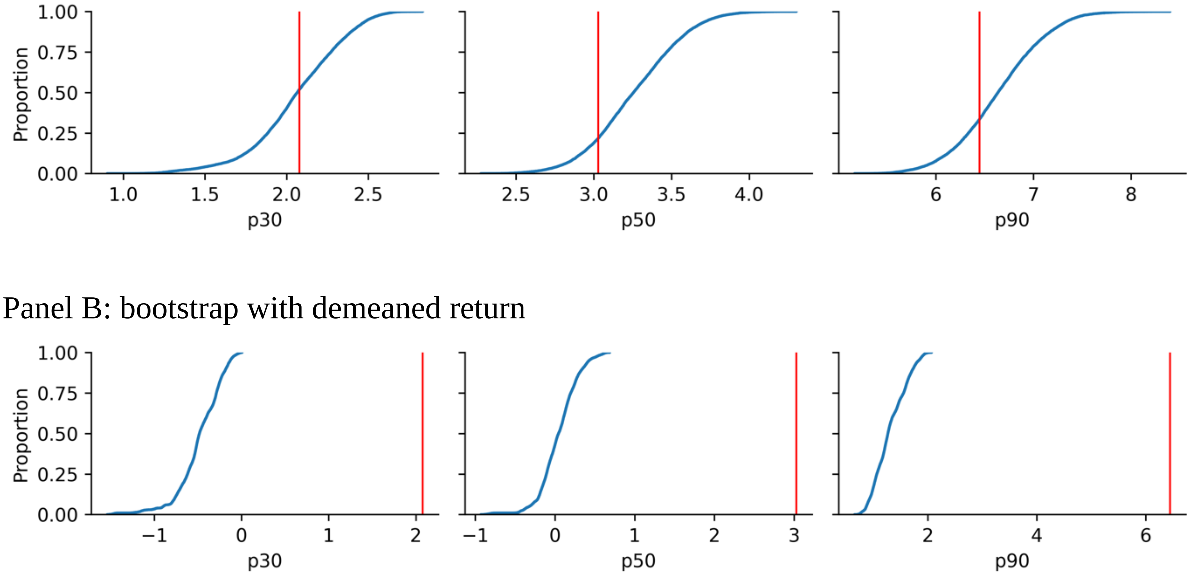


Figure 3. Empirical cumulative distribution of bootstrapped percentiles

The figure plots the empirical cumulative distribution function (ECDF) of cross-sectional percentiles (30%, 50% and 90%) based on two-step bootstrapping. In each bootstrap, the monthly long-short returns are computed using the average across specifications and then the cross-sectional percentiles are calculated. There are 10,000 simulations in total. In Panel A, raw returns are used and Panel B uses demeaned returns. The red vertical line is the actual cross-sectional percentile.



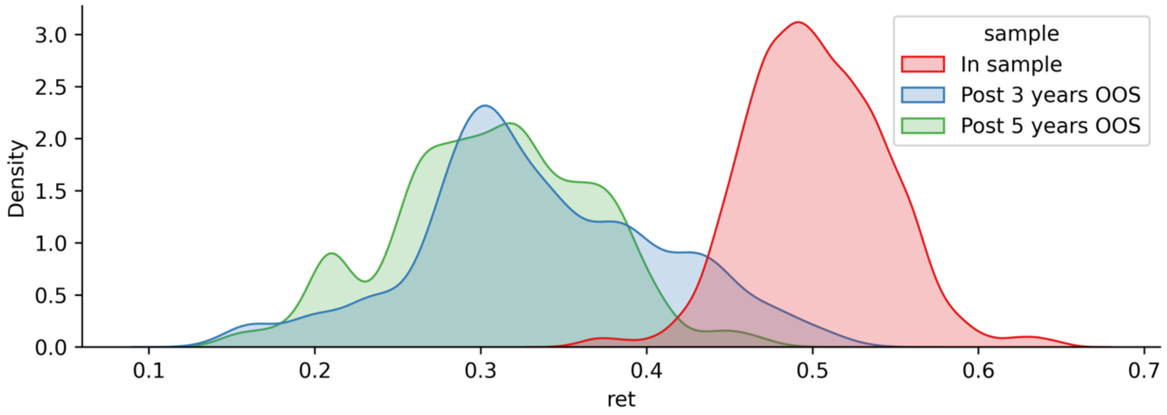
Panel A: bootstrap with raw return

Panel B: bootstrap with demeaned return

Figure 4. Out-of-sample performance

The figure provides distribution of bootstrapped cross-sectional mean returns of 173 anomalies in sample and out-of-sample. In sample is the time period that is used in the original paper. Out-of-samples include post 3 (5) years and pre 3 (5) years. Post samples use data from the end of date used in the original paper up to 3 or 5 years. Pre samples use the data that is from 3 or 5 years before the start of date used in the original paper. We run two-stage bootstrap in out-of-sample and in-sample periods. In each simulation, we calculate the cross-sectional mean of all anomalies and obtain the 10,000 averages.

Panel A: post 3- and 5-years out-of-sample



Panel B: pre 3- and 5-years out-of-sample

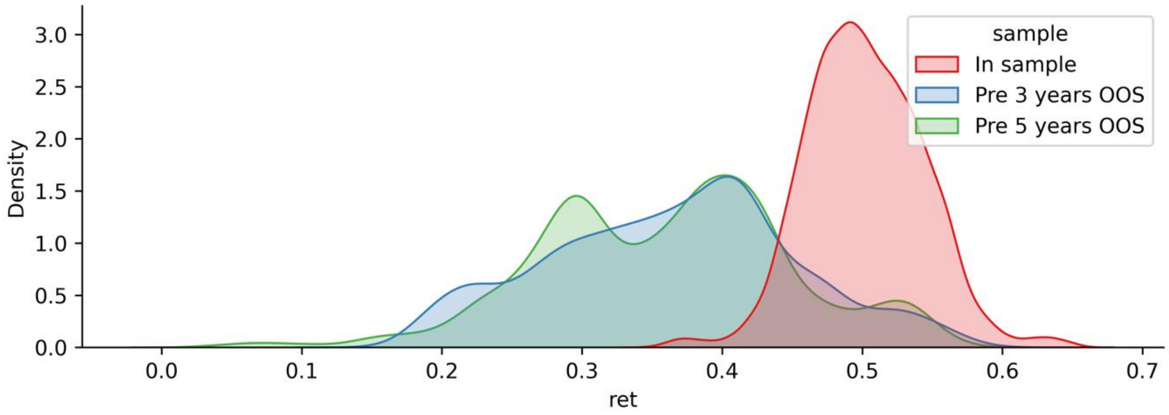
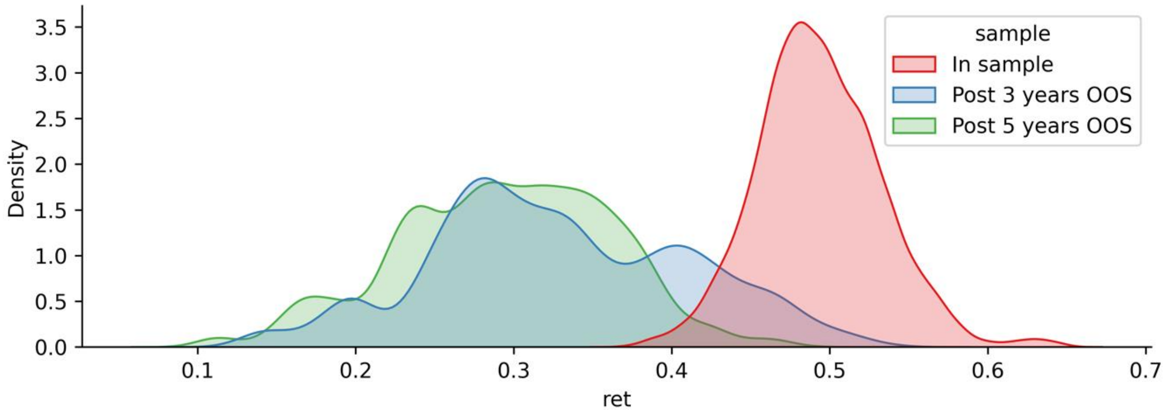


Figure 5. Post sample performance based on publication dates
The figure provides distribution of bootstrapped cross-sectional mean returns of 173 anomalies in sample and out-of-sample. In sample is the time period that used in the original paper. Out-of-samples include post 3- and 5-year. Post samples use data from the end of date used in the original paper up to 3 or 5 years. We run two-stage bootstrap in out-of-sample and in-sample periods. In each simulation, we calculate the cross-sectional mean of all anomalies and obtain the 10,000 averages. Panel A plots the distribution based on anomalies published after 1993 and Panel B plots the distribution based on anomalies published before 1993.

Panel A: anomalies published after 1993



Panel B: anomalies published before 1993

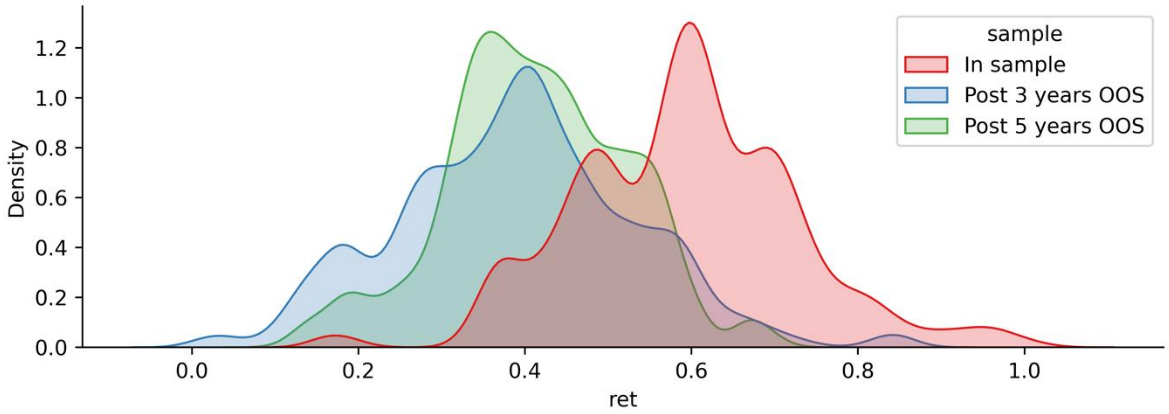


Figure 6. Histogram of 173 anomalies

The figure plots histogram of *t*-values in four settings: *t*-values estimated using OP methods, average *t*-value of anomaly from 10,000 bootstrapped samples based on the average aggregate, 10% anomalies select 70% percentiles and 20% anomalies select 60% percentiles. For the latter 2 settings, in each simulation, we randomly choose 10% (20%) anomalies which take the 70% (60%) percentile of the *t*-value across specifications rather than using the average.

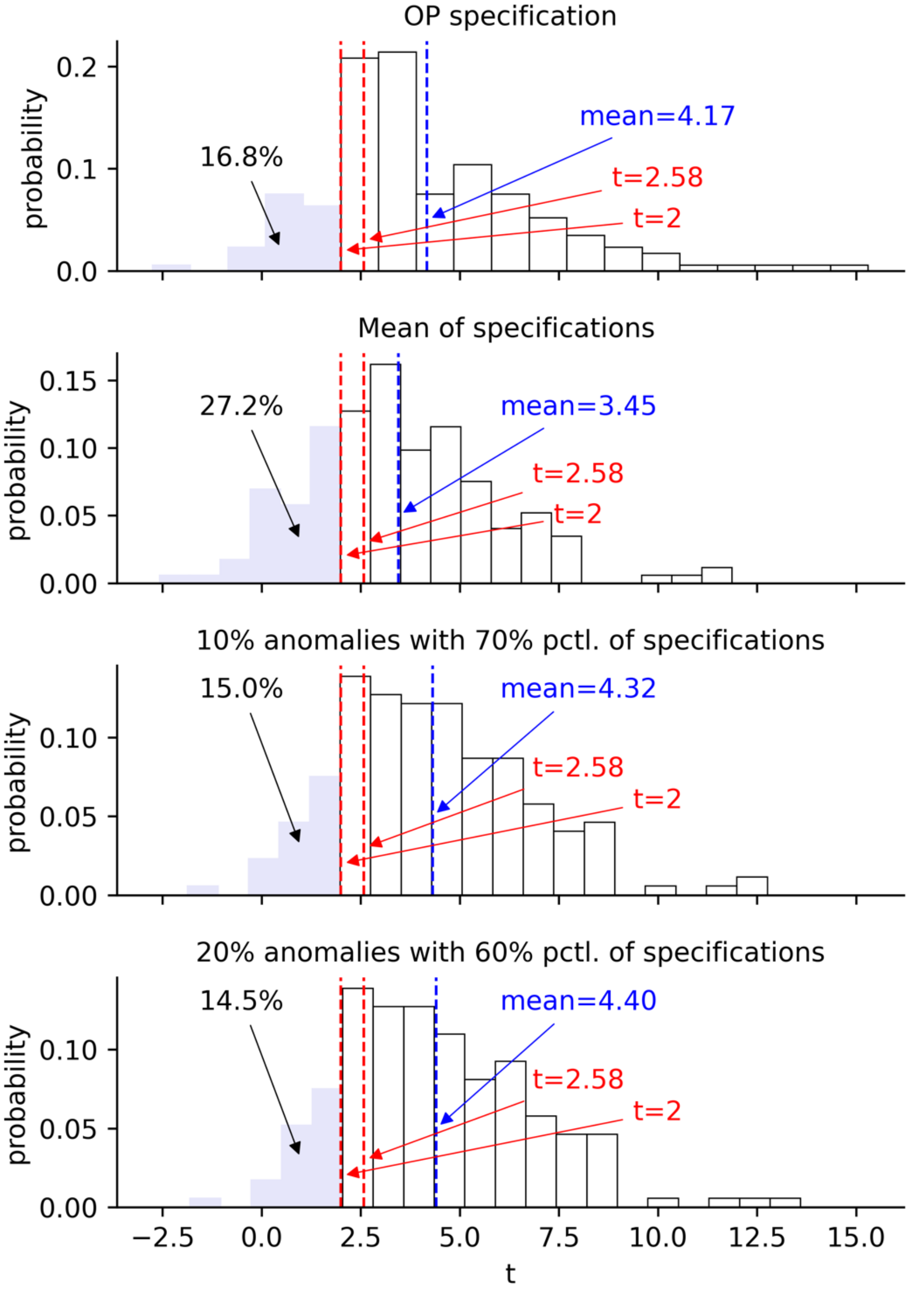
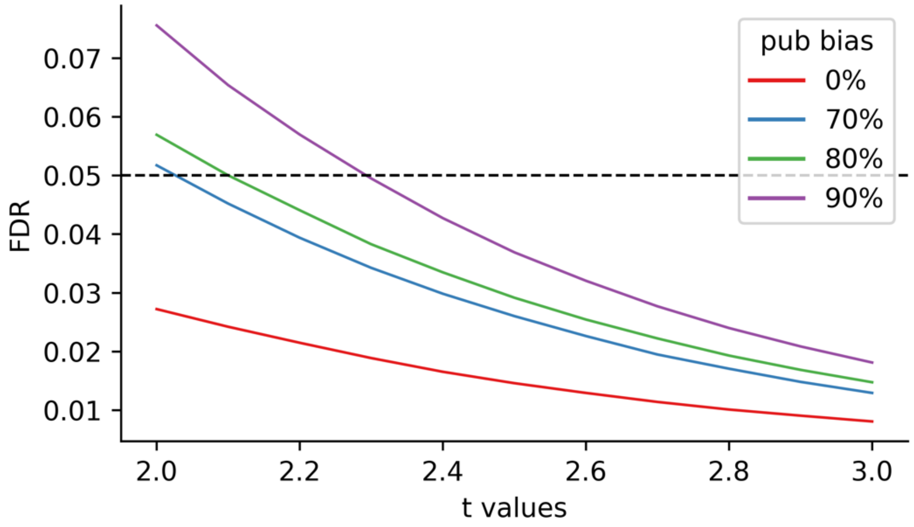


Figure 7. False discovery rates

The figure plots the false discovery rates (FDR) when the *t*-value cut-off between 2 and 3 under different extents of publications bias. Panel A (B) assume 70% (60%) anomalies are true, and we demean the returns of anomalies whose *t*-values are below 30% (40%) percentiles. In each bootstrapped sample, we calculate FDR as the number of significant findings that are false anomalies divided by the total number of significant findings. The significance is decided by *t*-values from 2 to 3. Publication bias is determined by how many insignificant specifications are dropped in each simulation.

Panel A: 70% anomalies are true



Panel B: 60% anomalies are true

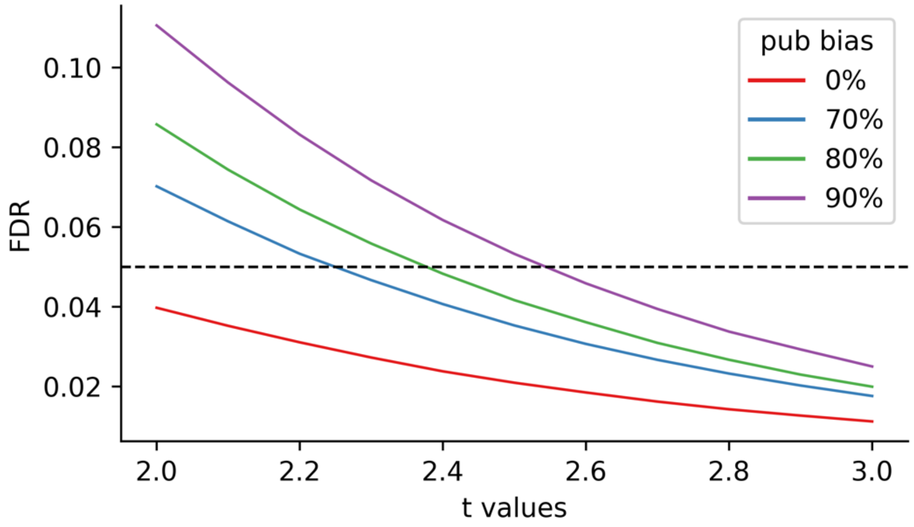


Table 1. Average anomaly returns and permutation test

The table reports average long-short return of the 173 anomalies and the percentage of rejection of null hypothesis using permutation test. Monthly long-short returns are first aggregated by either using the average or median across all specifications. The anomaly return is the time-series average of the monthly aggregate returns. Then we compute the cross-sectional average of all anomaly returns, ret(avg.) and ret (median). Column of ret (OP) is the average returns of all anomalies that are constructed following the methods in the original paper. Two permutation tests include hypothesis of same mean across specifications and hypothesis of same distribution of aggregate returns and OP returns. We also apply two grouping methods based on the explanations of anomaly and the data sources used to construct the anomaly.

	ret (avg.)	ret (median)	ret (OP)	pct. rejection H_0: same sample mean across specifications	pct. rejection H_0: same distribution of true (avg.) and OP	pct. rejection H_0: same distribution of true (median) and OP
All	0.36	0.35	0.50	55.5%	13.9%	17.3%
Agnostic	0.39	0.38	0.55	62.2%	13.3%	15.6%
Mispricing	0.38	0.37	0.52	56.0%	15.4%	18.7%
Risk	0.32	0.30	0.44	48.5%	12.1%	18.2%
Accounting	0.32	0.31	0.44	62.1%	16.1%	20.7%
Other	0.35	0.35	0.47	24.0%	4.0%	8.0%
Trading	0.42	0.41	0.60	59.0%	14.8%	16.4%

Table 2. Detailed results for selected predictors

This table reports returns, t values and permutation tests for 6 selected anomalies. Monthly long-short returns are first aggregated by either using the average or median across all specifications. The anomaly return is the time-series average of the monthly aggregate returns and t value for each anomaly is calculated. OP returns and t values of the 6 anomalies are also reported. The first two columns reports the results of permutation tests which are used to test hypothesis of same mean across specifications. F-value and the corresponding *p*-value are reported. The remaining columns reports the results of permutation tests which are used to test hypothesis of same distribution of aggregate (average or median) returns and OP returns.

			Average agg.				Median agg.				OP	
predictor	F	p (perm.)	Ret	t	OP diff	p (perm.)	Ret	t	OP diff	p (perm.)	Ret	t
size	4.71	0.0010	0.66	4.02	0.27	0.1918	0.52	3.29	0.12	0.5465	0.39	3.24
BMdec	1.83	0.0010	0.46	4.12	-0.25	0.1209	0.45	4.08	-0.26	0.1079	0.71	6.32
AssetGrowth	7.58	0.0010	0.44	5.46	-0.47	0.0020	0.40	5.08	-0.52	0.0010	0.91	7.32
GP	1.27	0.0460	0.47	5.25	0.06	0.6733	0.45	5.17	0.04	0.7622	0.40	3.87
RoE	0.74	0.9670	0.06	0.59	-0.15	0.3047	0.10	0.99	-0.11	0.4496	0.22	2.25
Mom12m	1.88	0.0010	0.82	4.38	0.07	0.8232	0.84	4.67	0.09	0.7592	0.75	3.01

Table 3. Replication rate across empirical decisions

This tables reports average returns, *t*-values and replication rate for different empirical decisions. For each anomaly, the monthly long-short returns are aggregated based on some certain specifications rather than all specifications and we take average of time-series returns as the anomaly return for that empirical decision. We report the results using the average aggregate in this table.

	Avg. ret	Avg. t	Pct. sig. (t>=2)	Pct. sig. (t>=3)
Delisting adj. ret = yes	0.36	3.41	0.72	0.51
Delisting adj. ret = no	0.36	3.40	0.72	0.51
Diff	0.00			
	(0.01)			
Ex. Price<=$5	0.33	3.21	0.69	0.47
No price filter	0.39	3.50	0.69	0.56
Diff	-0.06**			
	(-1.98)			
NYSE breakpoints	0.34	3.35	0.71	0.50
All stocks breakpoints	0.38	3.45	0.73	0.51
Diff	-0.03			
	(-1.25)			
VW	0.29	2.20	0.53	0.26
EW	0.43	4.28	0.80	0.61
Diff	-0.15***			
	(-4.96)			
Quintiles	0.32	3.24	0.71	0.50
Deciles	0.41	3.44	0.73	0.51
Diff	-0.09***			
	(-3.15)			
Ex. NASDAQ = yes	0.33	3.05	0.70	0.48
Ex. NASDAQ = no	0.37	3.45	0.71	0.51
Diff	-0.05*			
	(-1.76)			
Ex. Fin. = yes	0.37	3.38	0.71	0.50
Ex. Fin. = no	0.35	3.33	0.71	0.50
Diff	0.03			
	(0.94)			

Table 4. Regression: original-method anomaly returns and aggregate anomaly returns

The table reports the regression of *t*-values estimated from aggregate return on *t*-values estimated by method in the original paper. For each anomaly, the monthly long-short returns are aggregated based on some certain specifications rather than all specifications and we take average of time-series returns as the anomaly return for that empirical decision. We report the results using the average aggregate in this table.

	Coef	t	R2
Delisting adj. ret = yes	0.72***	18.27	0.73
Delisting adj. ret = no	0.72***	18.26	0.73
Ex. Price<=$5	0.64***	13.59	0.63
No price filter	0.76***	20.53	0.74
NYSE breakpoints	0.71***	18.27	0.73
All stocks breakpoints	0.72***	18.01	0.73
VW	0.39***	7.41	0.40
EW	0.99***	27.87	0.82
Quintiles	0.69***	19.15	0.73
Deciles	0.71***	15.38	0.67
Ex. NASDAQ = yes	0.63***	15.25	0.66
Ex. NASDAQ = no	0.74***	18.58	0.74
Ex. Fin. = yes	0.72***	17.96	0.73
Ex. Fin. = no	0.70***	17.99	0.72

Table 5. Bootstrapped replication rate

This table reports average return, *t*-value and replication rate from 10,000 two-stage bootstrapped samples. 95% confidence interval for each of the statistics is also reported. In the first stage of bootstrap, we randomly select months for all anomalies simultaneously. In the second stage of bootstrap, we randomly select specifications for all anomalies at the same time. Then we aggregate (average or median) long-short returns in each month using the bootstrapped sample and compute return and t value for each anomaly for that sample. We run first stage 100 times and we draw 100 times of the second stage for each run in the first stage. For each bootstrapped sample, we compute cross sectional mean of return, t value and replication rate (t cut-off at 2 and 3) of the 173 anomalies.

	Mean	95% CI lower	95% CI upper
Panel A: average agg.			
Ret	0.36	0.31	0.41
t	3.45	2.93	3.93
Pct. t>=2	70.5%	60.1%	77.5%
Pct. t>=3	54.4%	45.1%	63.6%
Panel B: median agg.			
Ret	0.35	0.29	0.41
t	3.43	2.71	4.14
Pct. t>=2	69.7%	58.4%	78.0%
Pct. t>=3	53.2%	40.5%	64.7%

Table 6. Bootstrapped replication rate: theory and data source groups

This table reports average return, *t*-value and replication rate from 10,000 two-step bootstrapped samples for different anomaly classifications. We define mispricing, risk and agnostic anomalies based on the explanations of anomaly using the classifications from Chen, Lopez-Lira and Zimmermann (2023). Using the classifications from Chen and Zimmermann (2022) based on the data sources, we also group anomalies to accounting, trading and other categories. In the first stage of bootstrap, we randomly select months for all anomalies simultaneously. In the second stage of bootstrap, we randomly select specifications for all anomalies at the same time. Then we aggregate (average or median) long-short returns in each month using the bootstrapped sample and compute return and *t*-value for each anomaly for that random sample. We run first stage 100 times and we draw 100 times of the second stage for each run in the first stage. For each bootstrapped sample, we compute cross sectional mean of return, *t*-value and replication rate (*t* cut-off at 2 and 3) of available anomalies in each classification.

	Average agg.			Median agg.		
Panel A: anomaly theory						
	Mispricing	Risk	Agnostic	Mispricing	Risk	Agnostic
Ret	0.38	0.32	0.40	0.38	0.30	0.39
t	3.59	3.29	3.62	3.61	3.15	3.61
Pct. t>=2	73.3%	68.9%	72.3%	72.6%	65.3%	72.9%
Pct. t>=3	57.2%	50.3%	56.4%	56.5%	47.9%	55.2%
Panel B: anomaly type						
	Accounting	Trading	Other	Accounting	Trading	Other
Ret	0.32	0.43	0.36	0.31	0.42	0.36
t	3.63	3.56	2.55	3.61	3.53	2.55
Pct. t>=2	74.7%	70.6%	55.9%	73.7%	69.3%	56.3%
Pct. t>=3	61.2%	52.2%	36.0%	60.2%	50.2%	36.6%

Table 7. Bootstrapped percentiles of *t*-values and *p*-values

This table presents actual percentiles of *t*-value, 95% confidence interval and *p*-values for each of percentile from 10,000 two-step bootstrapped samples. The two-step bootstrap procedure is as follow: in the first stage of bootstrap, we randomly select months for all anomalies simultaneously. In the second stage of bootstrap, we randomly select specifications for all anomalies at the same time. Then we aggregate (average or median) long-short returns in each month using the bootstrapped sample and compute return and t value for each anomaly for that random sample. We run first stage 100 times and we draw 100 times of the second stage for each run in the first stage. For 95% confidence interval, in each bootstrapped sample, we compute percentiles of t values of the 173 anomalies and obtain the distribution of different percentiles to compute the confidence interval for each percentile. For *p*-values, we demean monthly long-short anomaly returns first to assume zero anomaly return, and then perform the two-stage bootstrap. *P*-value (sim. pctl. ≥ act. pctl) measures how likely simulated percentiles are greater than the counterparty of actual percentiles. *P*-value (≥*t*) is how likely the 173 anomalies exceed each of the *t*-value.

Panel A: average aggregate

	Actual		95% CI of t		H0: Ret = 0	
	Ret	t	2.5%	97.5%	p-value (sim. pctl.≥act. pctl.)	p-value (≥t)
P0	-0.39	-2.57	-4.38	-1.02		
P10	0.06	0.45	-0.22	0.81		
P20	0.18	1.40	0.79	1.81		
P30	0.26	2.08	1.40	2.56	0.0000	0.0227
P40	0.30	2.70	2.02	3.26	0.0000	0.0037
P50	0.35	3.03	2.69	3.83	0.0000	0.0012
P60	0.39	3.73	3.29	4.54	0.0000	0.0000
P70	0.45	4.46	3.92	5.29	0.0000	0.0000
P80	0.49	5.17	4.68	6.21	0.0000	0.0000
P90	0.59	6.45	5.76	7.51	0.0000	0.0000
P100	1.49	11.72	10.71	14.25	0.0000	0.0000

Panel B: median aggregate

	Actual		95% CI of t		H0: Ret = 0	
	Ret	t	2.5%	97.5%	p-value (sim. pctl.≥act. pctl.)	p-value (≥t)
P0	-0.31	-2.15	-4.07	-1.09		
P10	0.06	0.55	-0.19	0.79		
P20	0.16	1.43	0.68	1.84		
P30	0.25	2.10	1.31	2.61	0.0000	0.0228
P40	0.30	2.68	1.91	3.32	0.0000	0.0053
P50	0.34	2.96	2.50	4.02	0.0000	0.0024
P60	0.37	3.67	3.02	4.72	0.0000	0.0001
P70	0.42	4.43	3.60	5.63	0.0000	0.0000
P80	0.48	5.10	4.28	6.61	0.0000	0.0000
P90	0.56	6.50	5.27	8.01	0.0000	0.0000
P100	1.49	12.49	10.45	15.73	0.0000	0.0000

Table 8. Binominal test: *p*-hacking

The table provides binominal test under the hypothesis that probability of above and below a threshold in a close window is equal. The thresholds are 2 and 3 and the widths away the threshold are 0.3 to 0.6 with a increment of 0.05. We report the number of anomaly, percentage of significant anomalies and *p*-value of observing as more extreme outcomes than the observed value. We run the test for OP anomalies in each window. The simulation results are based on the two-stage bootstrapped samples, and the results are the average of 10,000 simulations.

	OP			Simulation		
	Num. of anomaly	Pct. sig.	p-value	Num. of anomaly	Pct. sig.	p-value
2 ± 0.6	32	68.8%	0.0251	30.6	52.9%	0.4584
2 ± 0.55	27	63.0%	0.1239	28.0	52.6%	0.4750
2 ± 0.5	22	63.6%	0.1431	25.5	52.3%	0.4893
2 ± 0.45	19	68.4%	0.0835	22.9	52.0%	0.5042
2 ± 0.4	16	75.0%	0.0384	20.4	51.6%	0.5201
2 ± 0.35	14	71.4%	0.0898	17.9	51.3%	0.5348
2 ± 0.3	10	70.0%	0.1719	15.3	51.1%	0.5465
3 ± 0.6	49	49.0%	0.6123	34.8	50.0%	0.5509
3 ± 0.55	45	46.7%	0.7243	32.1	50.0%	0.5542
3 ± 0.5	44	47.7%	0.6742	29.3	50.0%	0.5558
3 ± 0.45	40	50.0%	0.5627	26.4	50.0%	0.5594
3 ± 0.4	33	54.5%	0.3642	23.5	49.9%	0.5638
3 ± 0.35	28	50.0%	0.5747	20.7	49.9%	0.5668
3 ± 0.3	25	52.0%	0.5000	17.8	49.9%	0.5708

Table 9. False discovery rate

The table reports false discovery rate (FDR) of published anomalies. Based on the bootstrapping results in Tables 5 and 7, we assume 70% of anomalies are true or 60% are true (if we apply lower bound of 95% confidence interval to be less confident about the published anomalies). The monthly long-short returns are demeaned if the *t*-values of anomalies are below 30% percentiles. Then we run the two-stage bootstrapping and select one specification randomly for each anomaly. For each bootstrap, we compute the FDR as the ratio between number of false anomalies that are significant and total number of significant anomalies. The reported FP, TP and FDR are the averaged number from 10,000 bootstrapped samples for number of false positive, number of true positive and false discovery rate. We use 2 and 3 as t value cut-offs. We repeat the procedure for simulation with publication bias. The publication bias is determined by how much proportion of insignificant specifications are dropped from each bootstrapped sample.

Panel A: 70% anomalies are true

	t-cutoff t=2			t-cutoff t=3		
pub bias	FP	TP	FDR	FP	TP	FDR
0%	2.7	95.8	2.7%	0.6	77.6	0.8%
70%	6.0	109.3	5.2%	1.1	86.7	1.3%
80%	6.8	112.3	5.7%	1.3	89.0	1.5%
90%	9.5	116.1	7.6%	1.7	91.3	1.8%

Panel B: 60% anomalies are true

	t-cutoff t=2			t-cutoff t=3		
pub bias	FP	TP	FDR	FP	TP	FDR
0%	3.6	85.8	4.0%	0.8	71.4	1.1%
70%	7.2	95.8	7.0%	1.4	78.6	1.8%
80%	9.2	98.4	8.6%	1.6	80.8	2.0%
90%	12.5	100.8	11.1%	2.1	82.6	2.5%

Appendix

Table A1. Average anomaly returns and permutation test for each anomaly

This table reports returns, t-values and permutation tests for each of the 173 anomalies. Monthly long-short returns are first aggregated by either using the average or median across all specifications. The anomaly return is the time-series average of the monthly aggregate returns and t value for each anomaly is calculated. OP returns and t values of the 6 anomalies are also reported. The first two columns reports the results of permutation tests which are used to test hypothesis of same mean across specifications. F-value and the corresponding *p*-value are reported. The remaining columns reports the results of permutation tests which are used to test hypothesis of same distribution of aggregate (average or median) returns and OP returns.

			Average agg.				Median agg.			
	F-stat	p (perm.)	ret	t	diff	p (perm.)	ret	t	diff	p (perm.)
AM	3.55	0.001	0.38	2.74	-0.11	0.584	0.29	2.16	-0.20	0.284
AOP	0.28	1.000	0.14	1.16	-0.03	0.836	0.13	1.11	-0.04	0.818
AbnormalAccruals	0.96	0.602	0.40	6.73	0.22	0.040	0.38	7.06	0.20	0.063
Accruals	1.55	0.001	0.37	5.43	-0.05	0.569	0.36	5.62	-0.06	0.495
AdExp	1.71	0.006	0.47	2.70	0.02	0.944	0.45	2.69	-0.01	0.980
AgeIPO	0.30	1.000	0.70	3.08	-0.05	0.862	0.77	3.28	0.01	0.956
AnalystRevision	1.99	0.001	0.56	6.34	-0.10	0.381	0.56	6.51	-0.10	0.356
AnalystValue	0.20	1.000	0.19	1.35	-0.05	0.824	0.21	1.53	-0.03	0.906
AnnouncementReturn	9.71	0.001	0.91	11.70	-0.24	0.024	0.90	12.49	-0.25	0.015
AssetGrowth	7.58	0.001	0.44	5.46	-0.47	0.002	0.40	5.08	-0.52	0.001
BM	4.66	0.001	0.49	2.87	-0.53	0.041	0.44	2.71	-0.59	0.026
BMdec	1.83	0.001	0.46	4.12	-0.25	0.121	0.45	4.08	-0.26	0.108
BPEBM	3.30	0.001	0.06	0.94	-0.09	0.335	0.08	1.41	-0.06	0.481
Beta	0.20	1.000	0.26	1.18	-0.11	0.715	0.22	1.03	-0.15	0.620
BetaFP	0.17	1.000	0.07	0.30	-0.02	0.958	0.03	0.13	-0.06	0.884
BetaLiquidityPS	0.13	1.000	0.04	0.39	-0.24	0.125	0.03	0.36	-0.25	0.117
BetaTailRisk	0.11	1.000	0.44	2.74	0.09	0.657	0.44	2.77	0.09	0.660
BidAskSpread	2.01	0.001	0.20	1.09	-0.45	0.157	0.09	0.55	-0.55	0.062
BookLeverage	0.60	1.000	-0.08	-0.93	-0.18	0.210	-0.06	-0.64	-0.15	0.285
BrandInvest	0.49	0.999	0.27	1.59	-0.15	0.620	0.13	0.75	-0.29	0.301
CBOperProf	2.15	0.001	0.57	5.54	0.04	0.842	0.58	6.02	0.05	0.784
CF	0.59	0.999	0.40	3.01	0.03	0.871	0.44	3.29	0.07	0.744
Cash	0.28	1.000	0.38	2.69	-0.13	0.602	0.37	2.69	-0.14	0.572
CashProd	1.60	0.001	0.21	1.92	-0.14	0.384	0.19	1.77	-0.15	0.344
ChAssetTurnover	0.64	0.996	0.26	4.50	0.09	0.228	0.24	4.40	0.07	0.316
ChEQ	2.37	0.001	0.47	4.95	-0.05	0.715	0.40	4.46	-0.11	0.431
ChInv	2.87	0.001	0.50	7.78	-0.11	0.274	0.48	7.79	-0.13	0.191
ChInvIA	3.64	0.001	0.22	3.68	-0.12	0.141	0.19	3.38	-0.15	0.065
ChNNCOA	0.59	1.000	0.27	5.24	0.05	0.461	0.27	5.50	0.05	0.466
ChNWC	0.64	0.997	0.26	5.34	0.11	0.078	0.24	5.54	0.09	0.107
ChTax	5.81	0.001	0.59	6.53	-0.33	0.019	0.60	6.96	-0.31	0.026
ChangeInRecommendation	1.38	0.011	0.34	3.76	-0.23	0.055	0.35	3.93	-0.22	0.067
CompEquIss	2.59	0.001	0.57	4.33	0.20	0.244	0.47	3.75	0.10	0.532
CompositeDebtIssuance	2.68	0.001	0.22	4.12	-0.01	0.869	0.23	4.56	0.00	0.995
CoskewACX	0.45	1.000	0.29	2.62	-0.07	0.639	0.31	2.80	-0.06	0.694
Coskewness	0.60	1.000	0.17	2.03	0.10	0.396	0.18	2.09	0.11	0.370
CustomerMomentum	0.51	0.999	0.88	4.62	0.03	0.919	0.85	4.57	0.00	1.000
DelBreadth	1.38	0.010	0.29	1.85	-0.27	0.295	0.27	1.77	-0.28	0.264
DelCOA	4.48	0.001	0.38	5.12	-0.03	0.756	0.36	5.22	-0.04	0.651
DelCOL	3.25	0.001	0.06	0.87	-0.15	0.097	0.06	0.87	-0.16	0.075
DelDRC	0.10	1.000	0.27	1.08	0.10	0.778	0.27	1.10	0.10	0.790
DelEqu	3.33	0.001	0.41	4.12	-0.07	0.649	0.34	3.65	-0.13	0.368
DelFINL	5.09	0.001	0.36	7.60	-0.14	0.034	0.34	7.58	-0.15	0.017
DelNetFin	1.10	0.254	0.29	5.71	-0.03	0.687	0.29	6.01	-0.03	0.648

Continued on next page

			Avg.				Median			
	F-stat	p (perm.)	ret	t	diff	p (perm.)	ret	t	diff	p (perm.)
DivYieldST	1.21	0.161	0.30	6.02	-0.25	0.002	0.30	5.92	-0.25	0.002
DolVol	3.84	0.001	0.46	3.59	-0.28	0.203	0.34	2.85	-0.41	0.057
EBM	2.87	0.001	0.30	4.08	0.07	0.472	0.26	3.62	0.03	0.738
EP	0.34	1.000	0.36	3.40	0.05	0.723	0.38	3.57	0.06	0.649
EarnSupBig	0.50	1.000	0.37	3.43	0.03	0.873	0.37	3.48	0.03	0.840
EarningsConsistency	1.06	0.347	0.27	3.10	0.00	0.973	0.30	3.71	0.03	0.783
EarningsForecastDisparity	0.57	0.999	0.40	3.05	-0.04	0.817	0.36	2.83	-0.08	0.641
EarningsStreak	2.29	0.001	0.80	7.60	-0.18	0.189	0.78	7.95	-0.19	0.145
EarningsSurprise	9.91	0.001	0.41	6.13	-0.27	0.007	0.43	6.50	-0.25	0.008
EntMult	1.04	0.389	0.51	4.40	-0.18	0.295	0.50	4.49	-0.18	0.286
EquityDuration	1.00	0.472	0.48	3.96	0.04	0.823	0.53	4.42	0.09	0.627
ExclExp	0.37	1.000	0.14	2.22	-0.02	0.829	0.15	2.39	-0.02	0.835
FEPS	0.32	1.000	0.43	2.02	-0.39	0.208	0.47	2.22	-0.35	0.254
FR	0.32	1.000	-0.06	-0.52	0.05	0.788	-0.01	-0.05	0.10	0.572
FirmAge	0.49	1.000	0.03	0.35	0.07	0.493	0.02	0.33	0.07	0.502
FirmAgeMom	1.66	0.001	1.22	6.47	0.03	0.927	1.26	6.94	0.06	0.821
ForecastDispersion	0.56	1.000	0.28	1.43	-0.16	0.530	0.26	1.38	-0.18	0.481
Frontier	5.69	0.001	0.49	2.80	-0.87	0.004	0.45	2.64	-0.91	0.003
GP	1.27	0.046	0.47	5.25	0.06	0.673	0.45	5.17	0.04	0.762
GrAdExp	1.46	0.021	0.34	2.74	-0.16	0.375	0.35	2.95	-0.15	0.394
GrLTNOA	1.64	0.001	0.12	2.00	-0.06	0.474	0.09	1.55	-0.09	0.262
GrSaleToGrInv	0.50	1.000	0.33	6.15	0.09	0.212	0.33	6.47	0.09	0.202
GrSaleToGrOverhead	0.33	1.000	0.03	0.47	0.04	0.621	0.03	0.52	0.04	0.609
Herf	0.62	1.000	0.02	0.31	-0.04	0.673	-0.03	-0.45	-0.09	0.344
HerfAsset	0.50	1.000	0.03	0.37	-0.01	0.934	0.02	0.26	-0.02	0.865
HerfBE	0.90	0.764	0.00	0.02	-0.06	0.534	0.00	-0.02	-0.07	0.507
High52	2.00	0.001	-0.14	-0.64	-0.13	0.663	-0.01	-0.06	0.00	0.998
IdioRisk	2.56	0.001	0.26	1.47	-0.29	0.281	0.35	2.10	-0.20	0.441
IdioVol3F	2.36	0.001	0.25	1.39	-0.28	0.290	0.34	2.05	-0.19	0.456
IdioVolAHT	2.36	0.001	0.03	0.17	-0.16	0.571	0.11	0.61	-0.08	0.741
Illiquidity	3.35	0.001	0.53	3.64	0.13	0.483	0.42	3.09	0.02	0.923
IndMom	3.25	0.001	0.38	3.01	0.02	0.888	0.36	2.81	0.00	0.995
IndRetBig	1.31	0.027	1.49	11.72	0.06	0.736	1.49	11.47	0.05	0.752
IntMom	1.58	0.001	0.53	3.75	-0.48	0.048	0.55	4.02	-0.46	0.060
IntanBM	1.32	0.023	0.38	2.08	0.10	0.665	0.43	2.36	0.15	0.530
IntanCFP	0.25	1.000	0.34	2.32	0.02	0.925	0.35	2.47	0.03	0.893
IntanEP	0.21	1.000	0.22	1.51	-0.08	0.619	0.24	1.74	-0.06	0.732
IntanSP	1.38	0.011	0.21	1.28	-0.21	0.336	0.16	1.06	-0.25	0.230
InvGrowth	2.98	0.001	0.45	6.08	-0.21	0.074	0.42	5.95	-0.24	0.039
InvestPPEInv	6.82	0.001	0.50	7.00	-0.03	0.747	0.49	6.84	-0.05	0.590
Investment	1.83	0.001	0.34	5.70	0.19	0.085	0.31	5.97	0.17	0.112
LRreversal	2.20	0.001	0.43	2.89	-0.26	0.262	0.36	2.55	-0.33	0.157
Leverage	2.57	0.001	0.36	2.79	-0.03	0.879	0.32	2.53	-0.07	0.681
MRreversal	1.88	0.001	0.46	3.78	0.03	0.883	0.42	3.56	-0.02	0.915

Continued on next page

			Avg.				Median			
	F-stat	p (perm.)	ret	t	diff	p (perm.)	ret	t	diff	p (perm.)
MS	0.14	1.000	0.41	4.29	-0.51	0.019	0.40	4.17	-0.52	0.018
MaxRet	1.17	0.122	0.32	1.94	-0.30	0.263	0.39	2.50	-0.23	0.395
MeanRankRevGrowth	1.98	0.001	0.11	1.32	-0.11	0.297	0.10	1.23	-0.12	0.242
Mom12m	1.88	0.001	0.82	4.38	0.07	0.823	0.84	4.67	0.09	0.759
Mom12mOffSeason	1.83	0.001	0.65	3.72	-0.22	0.449	0.69	4.06	-0.18	0.509
Mom6m	2.82	0.001	0.44	2.51	-0.10	0.698	0.49	2.96	-0.05	0.848
Mom6mJunk	1.02	0.418	0.45	1.74	-0.52	0.202	0.51	2.03	-0.46	0.242
MomOffSeason	4.40	0.001	0.63	5.19	-0.42	0.045	0.59	5.10	-0.46	0.030
MomOffSeason06YrPlus	1.38	0.013	0.39	5.07	-0.26	0.058	0.36	5.01	-0.28	0.041
MomOffSeason11YrPlus	0.49	1.000	0.18	2.75	-0.11	0.361	0.16	2.62	-0.13	0.263
MomOffSeason16YrPlus	0.24	1.000	0.17	2.46	-0.12	0.278	0.16	2.44	-0.13	0.235
MomSeason	1.16	0.151	0.50	5.82	-0.22	0.126	0.48	5.77	-0.24	0.084
MomSeason06YrPlus	1.53	0.001	0.55	7.53	-0.12	0.331	0.51	7.29	-0.15	0.194
MomSeason11YrPlus	1.10	0.223	0.46	7.46	-0.08	0.373	0.45	7.66	-0.09	0.337
MomSeason16YrPlus	0.63	1.000	0.34	5.43	-0.17	0.072	0.31	5.44	-0.19	0.036
MomSeasonShort	0.51	1.000	0.56	5.25	-0.29	0.091	0.55	5.35	-0.30	0.077
MomVol	0.21	1.000	0.38	2.07	-0.59	0.059	0.38	2.09	-0.60	0.053
NOA	2.75	0.001	0.55	6.99	-0.27	0.063	0.48	6.37	-0.33	0.019
NetDebtFinance	3.89	0.001	0.35	5.95	-0.31	0.001	0.34	6.17	-0.33	0.001
NetDebtPrice	0.67	0.994	0.06	0.50	-0.43	0.021	0.10	0.85	-0.39	0.030
NetEquityFinance	2.86	0.001	0.50	4.45	-0.44	0.026	0.54	4.98	-0.40	0.044
NetPayoutYield	1.16	0.140	0.31	2.58	-0.38	0.040	0.32	2.70	-0.37	0.044
NumEarnIncrease	4.03	0.001	0.29	5.21	-0.15	0.051	0.27	4.82	-0.17	0.035
OPLeverage	0.82	0.889	0.36	3.99	-0.04	0.788	0.33	3.69	-0.07	0.622
OperProf	0.83	0.884	0.33	2.98	-0.11	0.512	0.36	3.19	-0.08	0.588
OperProfRD	1.71	0.001	0.37	2.88	-0.05	0.786	0.43	3.56	0.01	0.961
OptionVolume1	0.37	1.000	0.31	1.66	-0.09	0.755	0.29	1.62	-0.10	0.727
OptionVolume2	0.18	1.000	0.28	2.61	0.00	1.000	0.23	2.33	-0.05	0.820
OrderBacklog	0.12	1.000	-0.07	-0.59	-0.15	0.308	-0.06	-0.48	-0.13	0.364
OrderBacklogChg	1.97	0.001	0.03	0.28	-0.33	0.062	0.03	0.24	-0.34	0.053
OrgCap	1.43	0.003	0.45	7.29	0.07	0.546	0.39	6.52	0.01	0.897
PS	1.14	0.169	0.45	2.64	-0.36	0.268	0.49	2.95	-0.32	0.331
PayoutYield	0.27	1.000	0.12	1.16	-0.13	0.375	0.13	1.25	-0.12	0.398
PctAcc	2.22	0.001	0.30	3.57	-0.11	0.370	0.33	4.19	-0.08	0.500
PctTotAcc	1.66	0.001	0.22	2.96	-0.16	0.144	0.22	3.28	-0.15	0.151
PredictedFE	0.21	1.000	-0.02	-0.12	-0.06	0.826	0.00	-0.02	-0.04	0.874
PriceDelayRsq	2.36	0.001	0.10	0.96	-0.43	0.027	0.03	0.26	-0.50	0.007
PriceDelaySlope	0.78	0.946	0.09	1.22	-0.12	0.294	0.08	1.15	-0.12	0.265
PriceDelayTstat	0.22	1.000	-0.02	-0.34	-0.04	0.667	-0.01	-0.19	-0.03	0.749
RD	3.04	0.001	0.49	3.81	-0.36	0.074	0.42	3.42	-0.44	0.034
RDAbility	0.67	0.993	0.23	1.42	0.10	0.641	0.22	1.36	0.08	0.662
RDS	0.48	1.000	0.23	2.70	-0.06	0.650	0.19	2.34	-0.10	0.474
RDcap	0.11	1.000	0.41	2.22	0.03	0.900	0.38	2.10	0.01	0.981
REV6	0.70	0.982	0.36	1.30	-0.31	0.400	0.42	1.64	-0.25	0.473

Continued on next page

			Avg.				Median			
	F-stat	p (perm.)	ret	t	diff	p (perm.)	ret	t	diff	p (perm.)
RIO_MB	0.47	0.999	0.46	2.80	-0.23	0.311	0.42	2.63	-0.27	0.244
RIO_Turnover	0.08	1.000	0.28	1.71	-0.08	0.736	0.27	1.63	-0.09	0.715
RIO_Volatility	0.20	1.000	0.38	1.74	-0.58	0.100	0.37	1.68	-0.59	0.101
ResidualMomentum	1.42	0.004	0.59	6.68	-0.24	0.086	0.57	6.55	-0.26	0.063
ReturnSkew	11.31	0.001	0.22	4.61	-0.22	0.008	0.24	5.09	-0.20	0.011
ReturnSkew3F	10.56	0.001	0.11	2.71	-0.23	0.001	0.13	3.10	-0.22	0.003
RevenueSurprise	7.74	0.001	0.27	4.23	-0.31	0.001	0.27	4.26	-0.32	0.001
RoE	0.74	0.967	0.06	0.59	-0.15	0.305	0.10	0.99	-0.11	0.450
SP	3.14	0.001	0.60	4.29	-0.16	0.398	0.57	4.17	-0.19	0.295
ShareIss1Y	1.52	0.002	0.39	6.07	-0.08	0.377	0.38	6.05	-0.09	0.359
ShareIss5Y	1.25	0.046	0.32	4.69	-0.10	0.318	0.30	4.49	-0.12	0.227
ShortInterest	2.86	0.001	0.41	2.91	-0.42	0.027	0.37	2.69	-0.45	0.019
SmileSlope	2.97	0.001	1.15	9.62	-0.05	0.762	1.13	9.87	-0.07	0.694
Tax	0.64	0.997	0.25	3.39	-0.10	0.338	0.30	4.15	-0.06	0.581
TotalAccruals	1.34	0.015	0.29	4.77	-0.01	0.932	0.25	4.28	-0.06	0.618
TrendFactor	4.23	0.001	1.39	10.71	-0.11	0.541	1.39	10.88	-0.11	0.550
VarCF	1.96	0.001	-0.39	-2.54	0.09	0.681	-0.31	-2.12	0.17	0.429
VolMkt	0.80	0.932	0.05	0.29	-0.21	0.337	0.06	0.33	-0.21	0.340
VolSD	1.43	0.011	0.33	3.23	0.08	0.595	0.29	2.96	0.04	0.781
VolumeTrend	1.39	0.010	0.44	4.64	-0.19	0.163	0.42	4.59	-0.21	0.133
XFIN	3.65	0.001	0.62	4.54	-0.43	0.078	0.63	4.85	-0.41	0.082
betaVIX	1.33	0.019	0.39	2.91	-0.18	0.429	0.32	2.52	-0.25	0.270
cfp	1.47	0.002	0.45	3.37	0.09	0.694	0.48	3.68	0.12	0.582
dNoa	5.88	0.001	0.60	7.98	-0.20	0.081	0.55	7.60	-0.25	0.037
fgr5yrLag	0.35	1.000	0.09	0.41	-0.06	0.838	0.12	0.55	-0.04	0.901
grcapx	1.92	0.001	0.34	5.21	-0.01	0.888	0.31	5.07	-0.04	0.629
grcapx3y	2.21	0.001	0.34	5.12	-0.03	0.742	0.34	5.44	-0.04	0.726
hire	5.17	0.001	0.28	3.17	-0.18	0.129	0.23	2.75	-0.22	0.050
price	3.19	0.001	0.49	2.44	-0.31	0.320	0.35	1.89	-0.45	0.141
realestate	0.33	1.000	0.18	2.72	-0.07	0.586	0.14	2.30	-0.11	0.392
retConglomerate	3.62	0.001	0.71	4.50	-0.45	0.046	0.66	4.18	-0.50	0.031
roaq	2.29	0.001	0.44	2.97	-0.74	0.003	0.51	3.60	-0.67	0.005
sfe	0.36	1.000	0.46	2.48	-0.11	0.738	0.52	2.87	-0.05	0.877
size	4.71	0.001	0.66	4.02	0.27	0.192	0.52	3.29	0.12	0.546
skew1	0.69	0.993	0.49	3.60	-0.02	0.931	0.50	3.91	-0.01	0.983
std_turn	0.43	1.000	0.47	2.76	0.07	0.779	0.47	2.82	0.07	0.767
strev	14.68	0.001	1.05	7.77	-1.67	0.001	0.99	7.61	-1.72	0.001
tang	1.18	0.090	0.13	1.37	-0.21	0.159	0.06	0.66	-0.27	0.053
zerotrade	1.90	0.001	0.36	2.61	-0.14	0.502	0.34	2.52	-0.16	0.451
zerotradeAlt1	1.88	0.001	0.29	2.05	-0.22	0.294	0.27	1.90	-0.24	0.252
zerotradeAlt12	2.54	0.001	0.39	3.03	-0.01	0.953	0.37	2.99	-0.03	0.868