MYTH BUSTED: STOCK RETURN ANOMALIES REVISITED

DISSECTING CROSS-SECTIONAL RETURN PREDICTABILITY IN ACCOUNTING-BASED ANOMALIES

ALMA FRIBERG

WILLIAM HU

Bachelor Thesis

Stockholm School of Economics

2021

Myth busted: Stock return anomalies revisited: Dissecting cross-sectional return predictability in accounting-based anomalies

Abstract:

Research has uncovered over 450 anomaly factors that exhibit stock return predictability. However, after anomalies are published and studied in successive literature, the return predictability often seems to attenuate or disappear. This raises the question of whether return predictability existed in the past, but have been arbitraged away, or whether published anomalies simply is an artifact of p-hacking. Using out-of-sample analysis, we study 21 eminent accounting-based stock return anomalies and show that a majority of the cross-sectional return predictability can be attributed to p-hacking.

Keywords:

Asset pricing, cross-sectional returns, anomalies, p-hacking, publication bias

Authors:

Alma Friberg (24384)
William Hu (24397)

Tutor:

Adam Altmejd, Researcher, Swedish House of Finance & Postdoc Fellow, Department of Finance

Examiner:

Adrien d'Avernas, Assistant Professor, Department of Finance

Bachelor Thesis
Bachelor Program in Business and Economics
Stockholm School of Economics

1. Introduction

1.1 Background and Relevance

A stock market anomaly is return predictability that is inconsistent with asset pricing models such as the Capital Asset Pricing Model (CAPM) or the Fama and French three factor model, implying an opportunity to earn anomalous stock returns. However, after anomalies are published and studied in successive literature, anomalies often seem to attenuate or disappear after their original sample. In finance research, competition for top journal space incentivizes p-hacking among researchers, which can cause reported results to exhibit low replicability in the future (Harvey et al., 2016). In a meta-study published in 2018, Hou et al. (2020) compiled the bulk of the published anomalies literature in finance and accounting by replicating 452 anomaly variables with a sample period from 1967 to 2016 and found that most anomalies failed to generate significant results. Mclean & Pontiff (2015) and Linnainmaa et al. (2018) performed meta-studies with separation between in-sample and out-of-sample time periods, and found that anomalies have significantly lower performance outside their original sample periods. This raises the question of whether return predictability existed in the past, but have been arbitraged away, or whether published anomalies simply is an artifact of p-hacking. To investigate the persistence and the mechanisms behind stock market anomalies, we replicate 21 eminent published accounting-based anomalies in four separate sample periods: Pre-sample, In-sample, Post-sample and also a Pre-publication period. These sub-sample periods allow us to discern the explanation for the existence of cross-sectional return predictability in anomalies, with the aim of determining whether published stock market anomalies has emerged due to p-hacking. Our study finds that 16 out of 21 anomalies have significant returns in the in-sample period, while only 2 out of 21 anomalies have significant returns in the post-sample period. Furthermore, after analyzing the results from our other sample periods and our z-score distribution, we conclude that a majority of the return predictability in our observed anomalies can be attributed to p-hacking.

Null hypothesis significance testing (NHST) is used for statistical inference in empirical finance. When an anomalous return has a p-value below the significance level of 0.05 it is usually interpreted as not being a statistical fluke. However, as shown by Gelman & Loken (2013), p-values are biased and many results identified may actually be due to chance. Inherently, there is a problem of multiple testing when multiple statistical tests are performed on the same dataset. Since hundreds of anomaly factors have been tested and not published, some significant factors are also bound to arise due to chance error. P-hacking disregards the multiple comparisons problem by performing many statistical tests on the data and cherry-picking significant results. Hence, academic journals contribute to p-hacking through their focus on publishing studies with significant results. However, the performance of anomalies identified through p-hacking are driven by sample error, and accordingly the return predictability does not hold up in out-of-sample tests. Out-of-sample denotes any sample period outside the original sample period. Therefore, performing out-of-

sample tests is a means of determining whether there is an indication of p-hacking in our observed accounting-based anomalies.

Our study is closely related to Linnainmaa et al. (2018), who replicate 36 accounting-based anomalies using Pre-sample, In-sample and Post-sample data. In-sample denotes the sample period in the original study, while Pre-sample and Post-sample denotes the sample periods before and after the original sample period, respectively. The sample frames allow them to discern between three competing explanations for the existence of anomalies based on: unmodeled risk, mispricing and p-hacking. If return predictability in a published study results solely from p-hacking, the predictability should disappear out-of-sample. Linnainmaa et al. (2018) find that the average returns and sharpe ratios of most anomalies decrease out-of-sample, while correlation among anomalies and volatilities increase. However, as found by McLean & Pontiff (2015), stock market anomalies are also less anomalous after being published, due to investors learning about mispricing from academic publications. Accordingly, this mechanism is a competing explanation for lower anomaly performance in the post-sample period. Therefore, we compute a pre-publication sample period in addition to the pre-sample and post-sample period, which denotes the period after the original sample period, but before the publication date of the study. This enables us further distinguish between mispricing and p-hacking effects by gauging anomaly performance in the period before the post-publication effects occurs.

The aim of a replication study is not to perfectly replicate the findings in each of the original papers. As mentioned by Mclean & Pontiff (2015), a replication that follows every detail would be impossible since CRSP data changes over time and papers often omit details about precise calculations. Rather, we implement the original definitions of the anomalies but use our own statistical testing framework, where the method or parameters in the replication procedure might differ from the original studies, to see if the hypothesis brought forth in the original paper holds. Our replication procedure is kept constant throughout all anomalies (see "methodology"). This also entails that anomalies which remain significant in our study in addition to the original study hold higher robustness, even if the anomaly turns out to be significant only in the in-sample period.

Furthermore, Brodeur et al. (2016) use a z-score plot to investigate the level of p-hacking and publication-bias in economics journals, where they plot the distribution of reported z-score from over 21,000 tests in a histogram. Their results indicate that there is an unnatural dip of reported z-scores right before significance at the 0.05 level. This can be an indication that tests that have shown close to significant results have been altered until they show a significant result. We construct a z-score plot similar to that of Brodeur et al. (2016) to investigate the z-score distribution of accounting-based anomalies. This allows us to gauge the level of p-hacking through other means than sub-period analysis as done by meta-studies such as Linnainmaa et al. (2018) and Mclean & Pontiff (2015) whose inferences were made from studying pre-sample, post-sample or post-publication sample periods. The

implications of the observed distribution is discussed by comparing with the distribution that would theoretically emerge absent p-hacking or publication bias.

1.2 Research Question

Our research aims to determine whether accounting-based return anomalies are persistent outside their original sample periods, and to examine the explanations for cross-sectional return predictability in the observed anomalies. Specifically, our study investigates whether the existence of published accounting-based anomalies can be attributed to p-hacking.

Therefore, our main research question is:

Is the cross-sectional return predictability in published accounting-based anomalies a result of p-hacking?

1.3 Literature Review and Contribution

Linnainmaa et.al. (2018) conduct a meta-study on a similar set of 36 accounting-based anomalies with pre-sample, in-sample and post-sample data. In addition to the sample periods of Linnainmaa et al., we add a pre-publication sample period in line with Mclean & Pontiff (2015) which allows us to further discern between competing explanations for the emergence of the anomalies. Furthermore, Linnainmaa et al. (2018) implement portfolio sorts with double sorted portfolios on size and the anomaly variable itself. We instead opt for a univariate sort, similar to Hou et al. (2020), based solely on the anomaly factor in order to focus on the performance in isolation. Finally, our paper adds to the post-sample period of Linnainmaa et al. (2018) since a longer post-sample period has accumulated since the implementation of their tests.

Hou et al. (2020) conducted a similar large scale replication with 452 anomaly variables, using univariate portfolio sort (and cross-sectional regressions). However, all anomalies were replicated with a sample period between 1967-2016 without sub-period analysis. Consequently, post-publication attenuation of anomaly performance could explain the failure to clear the 1.96 hurdle in the authors' replications. With their replication method, Hou et.al. found that 65% of the anomalies could not clear the single test hurdle of the absolute t-value of 1.96. Our contribution to Hou et al.'s study is therefore that we are able to discuss the competing explanations for the anomalies through out-of-sample analysis.

McLean & Pontiff (2015)'s meta study of 97 factors compare the predictor's return in-sample, post-publication and out-of-sample but pre-publication (which we simply call pre-publication in our study). They find that portfolio returns are 26% lower in the period after the original sample but before publication, which serves as an upper bound estimate of data mining effects. In the post-publication sample, they find that returns are 58% lower, and therefore attribute 32%(58% - 26%) to publication informed trading, i.e. that investors learn

about mispricing from academic publications and trade accordingly. Our study includes a pre-sample period in addition to the sample periods of McLean & Pontiff (201)5, which allows for further analysis of anomaly performance before the original sample period.

Our paper presents an additional analysis to further distinguish p-hacking from other competing explanations for declining anomaly performance. To gauge the publication bias and p-hacking, we look at the distribution of test statistics, as both publication bias (file drawer effect) and p-hacking induce certain patterns on the distribution of test-statistics and p-values (Harvey et al., 2016). Similar to Brodeur et al.(2016), we plot the z-scores found in the original studies to investigate whether the distribution show indications of p-hacking. This allows us to determine the extent of p-hacking through other means than out-of-sample analysis as done by Linnainmaa et al. (2018) and Mclean & Pontiff (2015).

2. Theory

This section provides the theoretical and mathematical framework of the asset pricing models that were used in the study, as well as the theoretical explanation for cross-sectional return anomalies. Additionally, it describes the statistical background of hypothesis tests and its implications.

Before we summarize and replicate asset price anomalies we here present the pricing models to which they are compared. An anomaly is simply an empirical observation where a systematic return can be observed that cannot be explained by standard asset pricing models. To judge whether an anomaly exist we thus first have to present the benchmark towards which they are identified.

1.1 Asset Pricing Models

We perform regressions against CAPM and Fama-French three-factor model (FF3) when computing the alpha of anomalies.

CAPM:

The capital asset pricing model sets out to price securities by risk and time value of money.

$$ER_i = R_f + \beta_i(ER_m - R_f)$$

where

$$ER_i = expected\ return\ of\ investment$$
$$R_f = risk - free\ rate$$
$$\beta_i = beta\ of\ investment$$
$$(ER_m - R_f) = market\ risk\ premium$$

The $\boldsymbol{Rf}$, risk- free rate, component accounts for the time value of money, while the $\boldsymbol{ER-Rf}$, market risk, component accounts for the risk of the asset. The β Beta represents the riskiness of the stock in relation to the market. In this way, the asset's volatility is taken into consideration in the pricing.

Fama and French Three Factor Model

The Fama and French Three Factor model is an extension to the Capital Asset Pricing model that adds size and value risk factors (Fama & French, 1992).

$$ER_i = R_f + \beta_1(ER_m - R_f) + \beta_2(SMB) + \beta_3(HML)$$

where

$$ER_i = expected\ return\ of\ investment$$
$$R_f = risk - free\ rate$$
$$\beta_{1,2,3} = factor\ coefficient$$
$$(ER_m - R_f) = market\ risk\ premium$$

$$SMB = size\ premium(small\ minus\ big)$$
$$HML = value\ premium(high\ minus\ low)$$

Through their research, Fama and French found that small-cap stocks and high book-to-market value stocks tend to outperform markets. By adding these factors to the CAPM, their three-factor model is able to explain 90% of diversification in portfolio returns, compared to CAPM's 70% (Fama & French, 1992). Apart from excess return on the market, which is also a factor in the CAPM, the three-factor model also contains the factors size of firm and book-to-market. The SMB factor accounts for the high returns from the small-cap firms, and the HML factor accounts for the high returns generated by stocks with a high book-to-market value.

1.2 Explanations for Cross Sectional Return-anomalies

Three main hypotheses are often mentioned as the cause for the existence of cross-sectional return anomalies (Linnainmaa, 2018). The first is unmodeled risk, which states that the anomalies come as an effect of the multidimensionality of stocks that cannot be reduced into a simple model, thus leading to misspecification. For instance, an anomaly might compensate for risk that CAPM or FF3 fails to account for. If return predictability exists as compensation for risk, the predictability should be persistent over time.
The implication that comes with this hypothesis is that the sample period should be irrelevant to the return predictability, as long as structural shifts in the risks that matter to investors do not occur. However, the past century has seen notable changes in cross-sections of corporate characteristics, meaning that a shift in significance of the anomaly may be a result of the shift in the underlying risk that drives the anomaly (Linnainmaa, 2018). This means that the choice of sample periods could matter if we have a non-diverse distribution of original sample periods.

The second hypothesis for why anomalies occur is mispricing. Mispricing exist due to the limits of arbitrage (Shleifer & Vishny, 1997), as well as investor irrationality (Linnainmaa, 2018). These factors cause asset prices to deviate from the price given by an asset pricing model. Over the past century, however, the possibility to exploit this arbitrage in mispricing has become easier (French, 2008). This is a result of both decreasing trading costs (Hasbrouck, 2009) and increased information availability due to digitalization and improved computing power. Mclean & Pontiff (2015) show that if return predictability reflects mispricing, publication will lead to investors learning about the mispricing, which will cause return predictability to decay. However, trading frictions will prevent anomalies to disappear completely.

The final hypothesis is p-hacking (data-snooping or data-mining), which would imply that the anomalies are significant by chance, also known as a type 1 error. Further explanation of hypothesis tests is presented in the following section.

1.3 Publication Bias, p-hacking, and hypothesis testing

Research can be considered more or less attractive based on whether or not it reports significant results (Szucks & Ionnaides, 2017). Fanelli (2013) finds that articles that are published with insignificant findings get fewer citations than those with significant findings. Hence, academic journals have a stronger incentive to publish articles that report significant findings. This is known as publication bias. P-hacking is broadly defined as the action of manipulating data in order to report statistically significant findings. Because of academic journals' preference for publishing articles which display statistically significant results, researchers find that when their research doesn't produce sub 0.05 p-values, a statistically significant result, it can be in their favor to strategically select or analyze in such a way that they get lower p-values. This can be based on decisions such as what data to include, what measures to study, and which interactions to measure etc. (Gelman & Loken, 2013). When using a 0.05 significance level in hypothesis testing, there is a 5% chance of finding a significant result when there is in fact not one. Although there are many forms of p-hacking, the form of p-hacking which is a common issue in cross-sectional literature is where published factors are actually drawn from multiple unpublished tests without accounting for multiple testing. Harvey (2017) illustrates in an empirical example that given a large enough choice set, dozens of long-short strategies based on the first three letters of stock tickers have significant t-statistics, which is an example of how p-hacking can work.

Type I and Type II errors in NHST

	Decision	
	$Fail\ to\ reject\ H_0$	$Reject\ H_0$
$H_0(true)$	Correct decision	Type I error ($\alpha\ error$)
$H_0(false)$	Type II error ($\beta\ error$)	Correct decision

Figure 1 demonstrates the possible outcomes in NHST based on the decision to fail to reject or reject H_0 and whether or not H_0 is true.

The value α represents the probability of a false rejection of the null hypothesis, also known as a type 1 error. Although this value is low, usually at 5%, it grows with the number of tests performed (Bickel & Docksum 1977). Due to the absence of a requirement for researchers to report all of their performed tests in their published work, there is no way to ensure that the significant results that appear from a study aren't a result of the performance of simply enough hypothesis tests.

1.4 Multiple testing adjustments

Another method of validating anomalies which are found through excessive data-mining or p-hacking in cross sectional research is to implement statistical adjustments that accounts for multiple testing. This method is an alternative validation method to out-of-sample testing presented previously. To account for multiple testing, the statistical literature often control for measures such as family-wise error rate (FWER) and false discovery rate (FDR). FWER is the probability of at least one type I error. FDP measures the expected proportion of false discoveries among all discoveries. There are many other measures or techniques which can be seen as extensions of the two aforementioned measures (Harvey et al., 2016).

Many statistical adjustment methods have been developed to control for both FWER and FDP. Harvey et al. (2016) use three different adjustments in their paper: Bonferroni's adjustment, Holm's adjustment and Benjamini, Hochberg and Yekutieli's adjustment (BHY). Since these multiple testing adjustments are all dependent upon the number of tests carried out, one needs to address the issue within cross sectional stock returns research - that some anomaly factors that are tested but are not made available to the public. Factors that have been tried and found insignificant are often discarded. Based on a simulation framework, Harvey et al. (2016) estimate that 71% of all tried factors are missing. The authors then establish a general t-statistic cutoff of 3.0 as a recommended cutoff for newly discovered anomaly factors. However, they mention that not necessarily all factors should be evaluated based on this cutoff. For instance, factors which are developed through theoretical principles should reasonably have a lower cutoff than a factor discovered through simply empirical testing. Nevertheless, we will use the 3.0 threshold to re-evaluate our anomalies as an additional computation to test our conclusions against.

3. Data and Methodology

3.1 Defining factor anomalies

We replicate 21 accounting-based anomalies whose factors can be further subcategorized into growth and investment, earnings quality, profitability and valuation. Table 1 lists these factors, along with their authors, original sample period, and the formula used to calculate the particular factor. (See "appendix" for a description of each anomaly).

Anomalies

Anomaly Factor	Authors	Original Sample Period	Formula
Growth and Investment			
Abnormal capital investment	Titman et al. (2004)	1973 - 1996	$\frac{capx_t}{revt_t} / (\frac{1}{3}\sum_{j=1}^{3}\left(\frac{capx_{t-j}}{revt_{t-j}}\right)$
Asset Growth	Cooper et al. (2008)	1968 - 2003	$\frac{at_t}{at_{t-1}} - 1$
Growth in Inventory	Thomas and Zhang (2002)	1970 - 1997	$\frac{invt_t}{(at_t + at_{t-1})/2}$
Growth in Sales minus Inventory	Abarbanell and Bushee (1998)	1974 - 1993	$\frac{(revt_t - \frac{revt_{t-1} + revt_{t-2}}{2})}{(revt_{t-1} + revt_{t-2})/2} - \frac{(invt_t - \frac{invt_{t-1} + invt_{t-2}}{2})}{(invt_{t-1} + invt_{t-2})/2}$
Investment growth rate	Xing (2008)	1964 - 2003	$\frac{capx_t}{capxt_{t-1}} - 1$
Investment-to-assets ratio	Lyandres et al. (2008)	1970 - 2005	$\frac{\Delta ppent_t + \Delta invt_t}{at_{t-1}}$
Investment-to-capital ratio	Xing (2008)	1964 - 2003	$\frac{capx_t}{ppent_{t-1}}$
Sustainable Growth	Lockwood and Prombutr (2010)	1964 - 2007	$\frac{ceq_t - ceq_{t-1}}{ceq_{t-1}}$
Earnings Quality			
Accruals	Sloan (1996)	1962 - 1991	$\frac{\Delta act_t - \Delta che_t - \Delta lct_t - \Delta dlc_t - \Delta txp_t - dp_t}{(at_{t-1} + at_t)/2}$
Net Operating Assets	Hirshleifer et al. (2004)	1964 - 2002	$\frac{(at_t - che_t) - (at_t - dlc_t - dltt_t - be_t)}{at_{t-1}}$
Net Working Capital Changes	Soliman (2008)	1984 - 2002	$\frac{\Delta(act_t - che_t) - \Delta(lct_t - dlc_t)}{at_{t-1}}$
Profitability			
Change in asset turnover	Soliman (2008)	1984 - 2002	$\Delta(\frac{revt_t}{at_t})$
Gross Profitability	Novy-Marx (2013)	1963 - 2010	$\frac{revt_t - cogs_t}{at_t}$
Operating Profitability	Fama and French (2015)	1963 - 2013	$\frac{revt_t - cogs_t - xsga_t - xint_t}{ceq_t}$
Profit margin	Soliman (2008)	1984 - 2002	$\frac{oiadp_t}{revt_t}$
ROA	Haugen and Baker (1996)	1979 - 1993	$\frac{ib_t}{at_t}$
ROE	Haugen and Baker (1996)	1979 - 1993	$\frac{ib_t}{ceq_t}$
Valuation			
Book-to-market ratio	Fama and French (1992)	1963 - 1990	$\frac{be_t}{mkvalt_t}$
Cash flow-to-price ratio	Lakonishok et al. (1994)	1968 - 1990	$\frac{ib_t + dp_t}{mkvalt_t}$
Enterprise Multiple	Loughran and Wellman (2011)	1963 - 2009	$\frac{mktvalt_t + dlc_t + dltt_t + pstkrv_t - che_t}{oibdp_t}$
Sales-to-price ratio	Barbee et al. (1996)	1979 - 1991	$\frac{revt_t}{mkvalt_t}$

Table 1 lists the individual anomalies that are replicated in our study. Column 1 lists each anomaly factor. Column 2 lists the author along with the year that their article was published. Column 3 lists the sample period in which the anomaly factor was tested in the original study. Column 4 lists the formula used in our study for each anomaly.

3.2 Data Sources

We obtain annual accounting data from the Compustat database, which provides bias-free coverage from 1963. Prior to 1963, there is a significantly sparser coverage for selected successful firms, since Compustat was established in 1962 and only backfilled information for selected firms. Therefore, we collected annual accounting data from 1963 - 2020 for all firms listed on NYSE, AMEX and NASDAQ. For the same period and stock universe, we collected stock returns data from CRSP. We also take delisting returns data from CRSP. Monthly Fama French factors and risk-free returns were obtained from Kenneth French's website. We only include common stocks (share codes 10 and 11). Unless stated otherwise in the original studies, rows with missing values (for variables used to construct the anomaly factors) are removed.

When studying the distribution of test statistics to identify p-hacking, the method used is similar to that of Brodeur et al. (2016), in which we aim to collect test-statistics that represent key hypotheses. T-statistics are collected from the original studies of the articles replicated by Linnainmaa (2018) and are solely from tables. T-statistics are collected from 22 original studies that find accounting-based anomalies. The scores collected are solely from tables and are recorded exactly as they are presented, i.e. we do not round up or down. Both negative and positive t-statistics are collected but are adjusted to their absolute value. Since the sample sizes in these studies are so large, we find that the conversion from t-statistics to z-scores results in such a minimal change that is too small to make a difference in the plot, we simply report the t-statistics as z-scores.

3.3 Sample Periods

We analyze the following sample periods encompassed by our data:

Sample Period Illustration

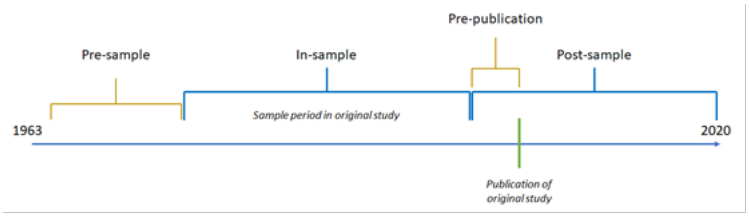

Figure 2 illustrates the sample periods used in out study on a timeline.

In-sample: the sample period used in the original study. This is computed for all anomalies.

Post-sample: the sample period occurring after the original sample period. This is computed for all anomalies.

Pre-sample: the sample period occurring before the original sample period. This sample period is computed for 9 anomalies with more than 7 years of pre-sample data (e.g. firms with original sample periods starting in 1964 will not have enough pre-sample data). We

implement this criterion to ensure that we have a sufficient sample size, since COMPUSTAT has lower data coverage for earlier years (lower amount of firms have sufficient accounting data).

Pre-publication out of sample: the sample period occurring between the end of the original sample period and the publishing of the article. This time frame usually spans between 2-6 years. This is computed for all anomalies. In this paper, we simply call this period "pre-publication".

We also present a distribution of our sample periods for each anomaly. Later in this paper, we refer to this figure when discussing sample selection sensitivity.

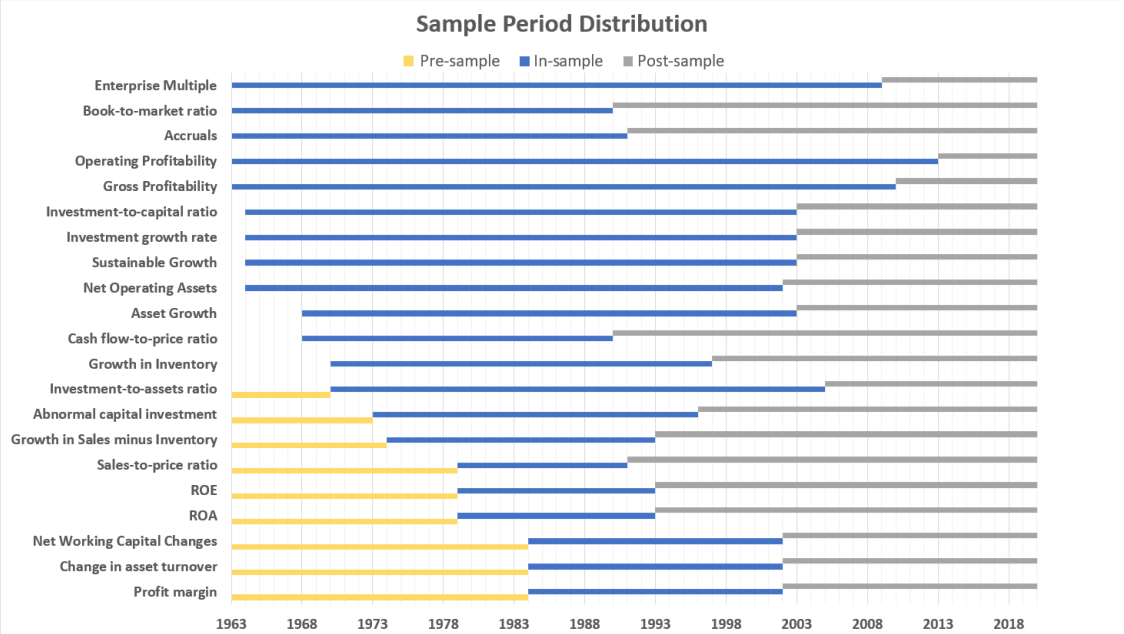

Figure 3. Plots the distribution of the original sample periods of each anomaly. The pre-publication sample period is not plotted in this graph, since it consists of the first 2-6 years in the post-sample period.

3.4 Methodology

To re-evaluate return anomalies we compute anomaly variables using the same formulas as the original studies. The definition of each anomaly is held constant throughout all sample periods to make the results comparable.

To evaluate the predictive power of an anomaly we use a portfolio sort approach, with stocks being sorted into annual quintiles (five portfolios) based on the anomaly variable. The quintile portfolios are then used to construct anomaly factors through high-minus-low approach. Thus, the return of the anomaly factor is the return of the highest quintile minus the return of the lowest quintile. The high and low labels are chosen based on the original study, where the stocks in the high portfolio earn higher returns than those in the low portfolios.

In computing the accounting-based anomalies and subsequent returns, we make the conventional assumption that accounting data is available six months after the fiscal year end date. We construct our stock anomalies with annual rebalancing at the end of June. In other words, stocks are sorted into portfolios in June year *t* with accounting information from the fiscal year that ended in year *t - 1*.

When forming portfolios, many studies use equal-weighted returns. However, we chose to compute value-weighted returns. Value weighted returns reflect the wealth effect experienced by investors (Fama, 1998). Furthermore, and more importantly, value-weighted returns help to control for microcaps. Microcaps account for about 60% of the total number of stocks, but only make up for 3% of the aggregate market capitalization of the NYSE-Amex-NASDAQ universe (Fama & French, 2008). High transaction costs cause anomalies present in microcaps to be less exploitable. Therefore, value-weighted returns is a more representative measure than equal-weighted returns.

The value weighted average monthly returns of each anomaly factor for each sub-period (in-sample, post-sample, pre-sample or pre-publication) are then tested against the null-hypothesis of zero return in a two-sided t-test. The anomaly factors are also regressed against CAPM and Fama French 3 factors to compute Alphas and corresponding t-statistics.

4. Results

4.1 Individual Anomalies

Table 2 presents the average monthly percentage return, as well as the CAPM and Fama-French 3-factor alphas. The average return, alphas and corresponding t-statistics is reported for the in-, post-, and pre-publication samples, as well as the pre- sample where applicable.

Average Returns, CAPM Alpha and FF3 Alpha for Individual Anomalies

	Average Return		CAPM		FF3	
Anomaly Factor	Avg.	t	Alpha	t	Alpha	t
EARNINGS QUALITY						
Accruals						
In	0.40	2.20*	0.51	3.04**	0.52	3.20**
Post	0.12	0.68	0.07	0.42	0.06	0.33
Pre	-	-	-	-	-	-
Pre-publication	0.40	1.19	0.67	1.98	0.04	0.16
Net Operating Assets						
In	0.44	3.47***	0.47	3.69***	0.45	3.51***
Post	0.32	1.95	0.18	1.12	0.22	1.39
Pre	-	-	-	-	-	-
Pre-publication	0.68	2.073*	0.41	1.24	0.15	0.50
Net Working Capital Changes						
In	0.66	3.34***	0.72	3.60***	0.75	3.71***
Post	0.12	0.09	0.05	0.52	0.04	0.22
Pre	0.59	3.34***	0.63	3.79***	0.58	3.53***
Pre-publication	-0.10	-0.36	-0.10	-0.38	-0.06	-0.20
PROFITABILITY						
Change in asset turnover						
In	0.39	2.27*	0.45	2.64**	0.34	1.96
Post	-0.19	-1.43	-0.23	-1.79	-0.22	-1.71
Pre	0.45	2.76**	0.46	2.89**	0.45	2.69**
Pre-publication	0.24	1.13	0.24	1.12	0.16	0.79
Gross Profitability						
In	0.28	1.65	0.36	2.14*	0.68	4.69***
Post	0.37	0.96	0.84	2.33*	0.23	0.79
Pre	-	-	-	-	-	-
Pre-publication	0.01	0.01	1.38	2.37*	1.08	2.47*
Operating Profitability						
In	0.41	2.03*	0.64	3.37***	0.75	4.40***
Post	0.16	0.63	0.26	1.04	0.21	0.88
Pre	-	-	-	-	-	-
Pre-publication	-0.23	-0.61	-0.07	-0.17	-0.15	-0.42
Profit margin						
In	1.28	2.67**	1.73	3.93***	1.19	3.74***
Post	-0.37	-1.00	0.12	0.35	0.20	0.73
Pre	-0.21	-0.85	-0.14	-0.65	0.42	2.49*
Pre-publication	-0.23	-0.30	-0.35	-0.63	-0.10	-0.21
ROA						
In	0.48	1.56	0.59	1.90	0.93	3.80***
Post	0.09	0.25	0.68	2.13*	0.64	2.55*
Pre	0.04	0.18	0.07	0.32	0.52	2.73**
Pre-publication	0.61	1.31	0.93	1.91	0.91	2.31*
ROE						
In	0.24	0.83	0.39	1.37	0.65	2.93**
Post	0.26	0.89	0.69	2.71**	0.67	3.05**
Pre	-0.45	-0.12	-0.19	-0.11	0.51	2.51*
Pre-publication	0.38	0.90	0.63	1.42	0.62	2.00*

VALUATION						
Book-to-market ratio						
In	0.60	2.77**	0.62	2.84**	-0.08	-0.79
Post	0.10	0.39	-0.04	-0.17	-0.24	-1.46
Pre	-	-	-	-	-	-
Pre-publication	0.73	0.84	0.96	1.02	-0.59	-1.18
Cash flow-to-price ratio						
In	1.12	4.16***	1.19	4.50***	0.52	2.67**
Post	0.25	0.72	0.78	2.48*	0.68	2.74**
Pre	-	-	-	-	-	-
Pre-publication	0.58	0.87	0.84	1.30	0.89	1.96
Enterprise Multiple						
In	0.24	1.16	0.17	0.81	0.00	-0.01
Post	0.11	0.35	-0.35	-1.25	-0.24	-0.98
Pre	-	-	-	-	-	-
Pre-publication	-0.16	-0.33	-0.30	-0.64	-0.17	-0.39
Sales-to-price ratio						
In	0.67	2.44*	0.71	2.58*	0.22	1.09
Post	0.22	1.01	0.29	1.29	0.13	0.90
Pre	0.69	2.20*	0.63	3.79***	0.58	3.53***
Pre-publication	0.39	0.94	0.61	1.45	-0.37	-1.43
GROWTH AND INVESTMENT						
Abnormal capital investment						
In	0.42	2.97**	0.39	2.71**	0.37	2.50*
Post	0.29	1.52	0.28	1.43	0.28	1.51
Pre	0.37	2.14*	0.37	2.13*	0.33	1.91
Pre-publication	0.56	1.35	0.63	1.53	0.70	1.70
Asset Growth						
In	0.49	3.36***	0.55	3.84***	0.28	2.14*
Post	-0.09	-0.55	-0.08	-0.47	0.04	0.27
Pre	-	-	-	-	-	-
Pre-publication	-0.28	-1.01	-0.28	-1.02	-0.41	-1.54
Growth in Inventory						
In	0.38	2.60**	0.47	3.37***	0.31	2.21*
Post	0.90	2.30*	0.95	2.63**	1.05	2.78**
Pre	-	-	-	-	-	-
Pre-publication	0.85	2.17*	0.91	2.55*	1.02	2.73*
Growth in Sales minus Inventory						
In	0.39	2.54**	0.41	2.69**	0.53	3.37***
Post	0.08	0.52	0.08	0.48	0.07	0.42
Pre	0.81	4.11***	0.80	4.22***	0.82	4.17***
Pre-publication	0.43	1.66	0.44	1.60	0.31	1.17
Investment growth rate						
In	0.46	3.41***	0.50	3.73***	0.31	2.37*
Post	0.01	0.04	0.09	0.40	0.15	0.72
Pre	-	-	-	-	-	-
Pre-publication	-0.64	-1.70	-0.70	-1.94	-0.76	-2.12
Investment-to-assets ratio						
In	0.50	4.03***	0.56	4.62***	0.33	2.90**
Post	-0.06	-0.31	-0.11	-0.60	0.14	0.84
Pre	0.60	1.83	0.62	1.90	0.18	0.63
Pre-publication	-0.06	-0.11	-0.27	-0.50	-0.04	-0.11
Investment-to-capital ratio						
In	0.45	1.98*	0.67	3.35***	0.22	1.34
Post	-0.25	-1.15	-0.24	-1.06	0.06	0.33
Pre	-	-	-	-	-	-
Pre-publication	0.40	1.20	0.39	1.33	0.29	1.29
Sustainable Growth						
In	0.32	2.54**	0.39	3.09**	0.11	0.93
Post	0.20	1.21	0.21	1.26	0.38	2.44*
Pre	-	-	-	-	-	-
Pre-publication	0.51	1.48	0.52	1.47	0.64	2.04*

Table 2 displays average monthly returns, CAPM Alpha and FF3 Alpha, and their respective t-statistics for individual anomalies. Asterisks after the t-statistics indicate the p-values: * = P < 0.05, ** = P< 0.01, *** = P<0.001.

In-sample, we find that 16 out of 21 anomaly factors earn returns that are statistically significant at the 5% level. 18 anomaly factors also have significant CAPM alphas, and 16

have significant three-factor model alphas. For some anomalies, such as Change in asset turnover or ROE, the difference in significance between returns and either CAPM alphas or three-factor model alphas are large compared to other anomalies. This has to do with how an anomaly covaries with the market and the FF3 factors. Linnainmaa (2018) explains that an anomaly that covaries negatively with the market and FF3 factors might exhibit low returns but considerably higher alphas, for instance.

In the post-sample period, only 2 out of 21 anomalies have significant returns at the 5% level. 5 anomalies have significant CAPM alphas, and 5 also have significant three-factor model alphas. Thus, we see a markedly weaker performance of anomalies in the post-sample period.

We compute pre-sample returns and alphas for 9 anomalies with more than 10 years of available pre-sample data, and also pre-publication returns and alphas for all anomalies. We find that 5 out of 9 anomalies have significant returns pre-sample. Furthermore, 5 and 7 have significant CAPM and three-factor model alphas in the pre-sample period respectively. In the pre-publication sample, only 2 out of 21 anomalies have significant returns, 3 have significant CAPM alphas, and 6 have significant three-factor model alphas. However, especially for the pre-sample and pre-publication periods, the statistical power for any one anomaly is limited due to the smaller sample size. Therefore, it is also preferable to look at the aggregate of all anomalies and compute averages, as in the section below (see "4.2 Average Anomalies").

Table 3 summarizes our results for individual anomalies:

Number of significant t-statistics by sample period

T-statistic of:	Pre-sample	In-sample	Pre-publication	Post-sample
Average return	5	16	2	2
CAPM alpha	5	18	3	5
FF3 alpha	7	16	6	5
Total anomalies	9	21	21	21

Table 3. The table shows the count of significant average returns, CAPM Alphas and FF3 Alphas by sample period. For In-sample, Post-sample and Pre-publication we computed t-statistics for all 21 anomalies. Pre-sample t-statistics are computed for 9 anomalies.

4.2 Average Anomalies

In this section we show results for the average of all observed anomalies. Average return, CAPM Alpha and FF3 Alpha averages are computed for each sample period.

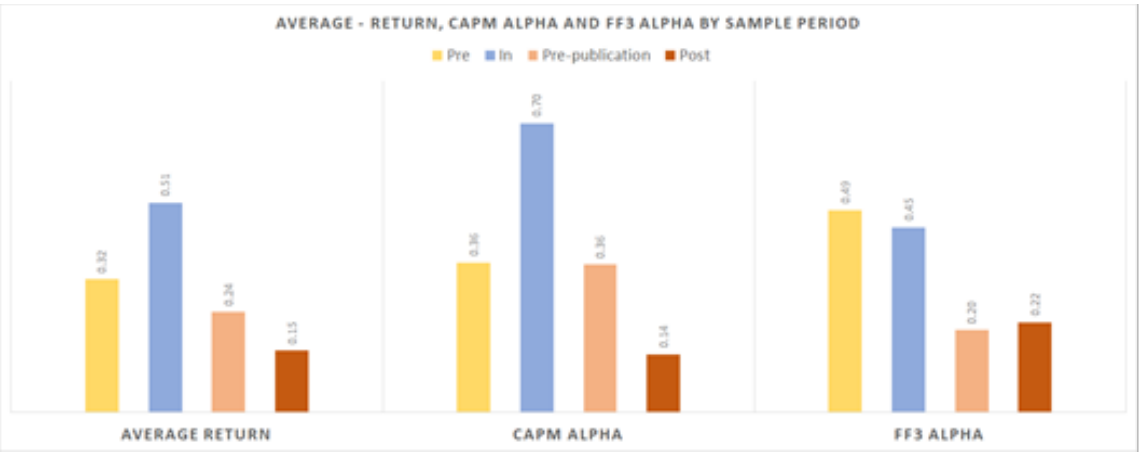

Figure 4 plots the average return, CAPM Alpha and FF3 Alpha by sample period.

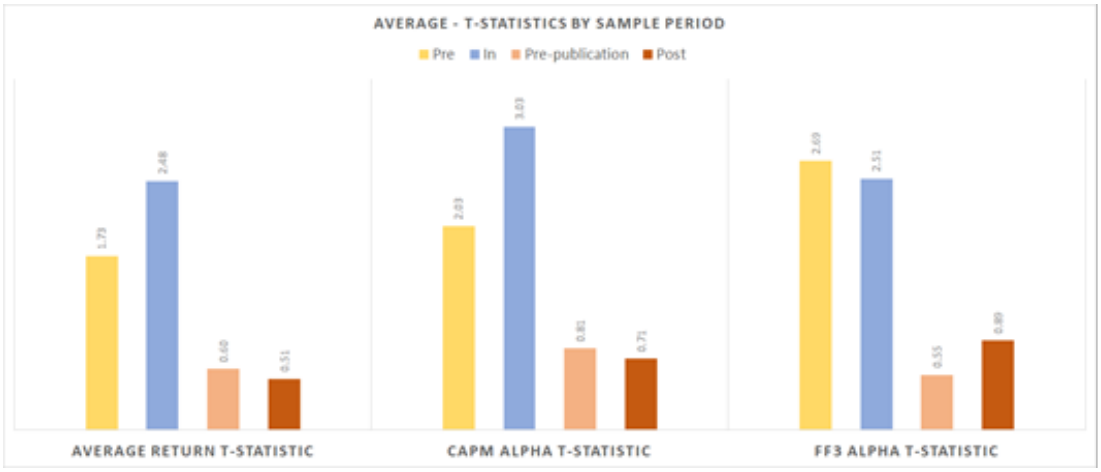

Figure 5 plots the t-statistics of average return, CAPM Alpha and FF3 Alpha by sample period.

Figure 4 shows that the average anomaly factor earns a monthly return of 0.51% during the in-sample period. In the post-sample period, the return decreases to 0.15% per month, while the pre-sample period indicates a slightly higher return of 0.32%. In the pre-publication sample we also see a significant depletion of returns, with 0.24% in monthly return. The anomalies also show a similar relationship in CAPM Alpha, with an in-sample CAPM Alpha of 0.70 and notably lower alphas in out-of-sample periods, with a post-sample alpha of 0.14. For FF3 Alpha, the pre-sample period exhibits a slightly higher alpha than the in-sample period.

In addition to the alpha, volatility should be taken into consideration when determining the attractiveness of an anomaly as an investment. Although our out-of-sample anomaly factor returns are lower, a decrease in volatility could counteract this. Sharpe ratio takes volatility into consideration through dividing an anomaly factor's average return by its volatility (standard deviation of returns). Figure 6 shows average sharpe ratios by sample period.

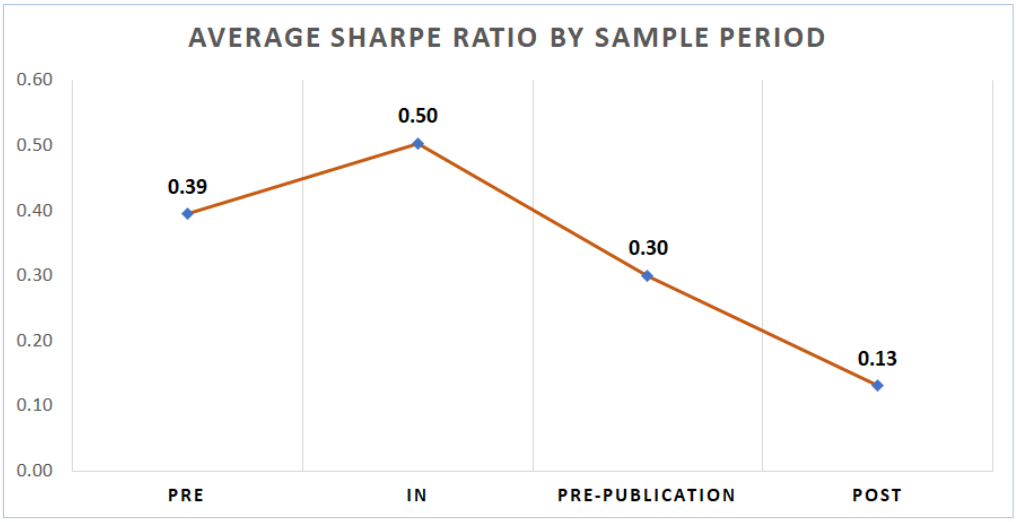

Figure 6. Average sharpe ratio by sample period. We have annualized the monthly sharpe ratios by multiplying with $\sqrt{12}$

We see that the average sharpe ratio is clearly higher in-sample than out-of-sample, which further strengthens our findings that out-of-sample anomaly factor performance is lower than in-sample. The figure also indicates a declining trend between pre-sample, pre-publication and post-sample periods.

It is also interesting to look at t-statistics grouped by anomaly category to investigate whether certain anomaly categories perform better.

Anomaly significance by category

	Pre			In		
Category	Average Return	CAPM Alpha	FF3 Alpha	Average Return	CAPM Alpha	FF3 Alpha
Earnings Quality	3.34	3.79	3.53	3.01	3.45	3.48
Growth and Investment	2.69	2.75	2.24	2.93	3.42	2.22
Profitability	0.49	0.61	2.60	1.53	2.52	3.59
Valuation	2.20	3.79	3.53	2.63	2.68	0.74
	Pre-publication			**Post**		
Category	Average Return	CAPM Alpha	FF3 Alpha	Average Return	CAPM Alpha	FF3 Alpha
Earnings Quality	0.97	0.95	0.15	0.90	0.69	0.65
Growth and Investment	0.63	0.63	0.64	0.45	0.51	1.16
Profitability	0.40	1.00	1.16	0.25	1.08	1.05
Valuation	0.58	0.78	-0.26	0.72	0.59	0.30

Table 4. Average return, CAPM Alpha and FF3 Alpha t-statistics by anomaly category and sample period.

Table 4 shows us that different anomaly categories perform better with regards to alphas or average returns. Profitability anomalies have lower performance than other categories with regards to average return, but still exhibit competitive CAPM Alpha and FF3 Alpha in most sample periods. Also, valuation anomalies have significantly lower FF3 Alpha compared to other anomalies in all periods except pre-sample. Later in this study, we will provide plausible explanations for these results.

6. Discussion

6.1 Competing explanations for the emergence of anomalies

Unmodeled risk

If the cross-sectional return predictability in anomalies is an effect of unmodeled risk, risk that is not accounted for by risk models such as CAPM or the three-factor model, then we expect the effect to be similar across time periods. That is, if return predictability exists as compensation for risk, the predictability should be persistent over time, as with the factors of CAPM or FF3 for instance. However, as mentioned previously, there is still the possibility that structural shifts in the risks that matter to investors cause the anomalies to perform differently across time periods.

Consider the case where all anomalies have the same in-sample periods (also pre- and post-sample), then changing markets conditions or investor behavior during a specific time period could possibly explain changes in return predictability throughout our sample periods. For instance, macroeconomic events or other events could create a structural break which coincides with the shift between In- and Post- sample periods, which causes anomaly performance to decline Post-sample. However, this would be mitigated by a diverse distribution of original sample periods (or In-sample periods), since a time specific change will not only be present in one type of sample period. In our sample period distribution graph, we see that although several anomalies have original sample periods starting in the 1960s, there are also many anomalies which start at significantly later time frames. Additionally, anomalies with longer sample periods are less sensitive to structural shifts occurring during that period. Therefore, we judge that our conclusions are not particularly sensitive to changes in market conditions during a specific period. Hence, the unmodeled risk hypothesis is a weak explanation for the non-persistent anomaly performance throughout our sample periods.

Mispricing

Under the mispricing hypothesis, the publication of an anomaly paper should cause sophisticated investors to learn about the mispricing and trade against it. Therefore, the cross-sectional return predictability should disappear or decay after publication. Accordingly, this is consistent with lower anomaly performance in the post-sample period. However, this does not explain the attenuation of anomaly returns between the in-sample period and pre-publication period which is a sample period before the publication date. Hence, mispricing alone cannot explain the decay of anomaly performance in the post-sample period. On the other hand, one should take into consideration that the release of research as working papers could contribute to informed trading before the publication date, which we will discuss later.

Similar to Mclean & Pontiff we estimate an upper bound for the mispricing effect. In this case, the performance decline between pre-publication and post-sample is attributed to mispricing effects. As presented in our average anomaly results, we find that anomaly returns

are 52.3% lower in the pre-publication sample compared to in-sample, in relative terms. Also, anomaly returns are 70.8% lower in the post-sample relative to the in-sample period. An upper bound estimate of the mispricing effect is therefore a 18.48% decline of in-sample returns (70.8% - 52.3%).

A part of the mispricing hypothesis also involves declining limits to arbitrage, implying that restrictions that prevent investors from exploiting mispricing opportunities are decreasing, which causes mispricing to be less prevalent over time. Therefore this theory is consistent with declining performance throughout the sample periods in chronological order: pre-sample, in-sample, pre-publication and post-sample. Evidently, our results implies that mispricing cannot explain our pre-sample performance which is lower than the in-sample performance. However, we do see a declining performance between the pre-sample, pre-publication and post-sample with regards to average returns, CAPM alpha but also Sharpe-ratio. This trend of declining performance is consistent with declining limits to arbitrage.

Mclean & Pontiff(2015) discusses the possibility that publication of academic research has no effect on return predictability, but that the decline in performance between the pre-publication and post-sample is explained by time trends such as declining limits to arbitrage. For instance, declining anomaly returns post-sample may simply be explained by lower trading costs and increase in hedge funds in later time periods. In order to investigate the possibility that their results reflect time effects and not a publication effect, the authors conduct regressions with a time variable. However, they find that the post-publication coefficient is still statistically significant, implying that publication of academic research does have a significant effect on the decay of anomaly returns.

P-hacking

P-hacking suggests that significant anomalies emerge by chance due to testing of multiple anomalies. The process of p-hacking involves testing numerous hypotheses using the same data set. Under the p-hacking hypothesis, the in-sample performance of anomalies are unique to that period and become insignificant out-of-sample. Therefore, our findings are consistent with p-hacking, where in-sample performance is markedly higher than out-of-sample performance.

As previously computed for mispricing, we can discuss an upper bound estimate of the p-hacking effect. The lower performance during the pre-publication period occurs before the publication date, so publication-informed trading (exploitation of mispricing) cannot explain the performance decrease between the pre-publication and the in-sample period. However, as previously mentioned, declining limits to arbitrage can explain lower anomaly returns throughout the sample periods in chronological order. Therefore, declining limits to arbitrage could play a part in the attenuation of anomaly performance between the in-sample and pre-publication periods. Nevertheless, it is unlikely for the entire decline of 52.3% to be explained by declining limits to arbitrage. Therefore, assuming that declining limits to

arbitrage have a minor effect on the decay of anomaly returns, a 52.3% depletion of in-sample returns is still an upper bound estimate of p-hacking effects. Note that this is an upper bound estimate, since research is often released as working papers sophisticated traders might be aware of the mispricing before publication, as mentioned by Mclean & Pontiff (2015). Some traders are likely to learn about the predictor before publication and cause the decay in the pre-publication period to be larger than the actual decay from p-hacking.

To further investigate the extend of p-hacking in the original articles that Linnainmaa (2018) replicate (from which we select our anomalies that we replicate), we also collect and plot the z-scores that 22 of those articles present in their tables, similar to the process of Brodeur et al. (2016) (as described under "data sources").

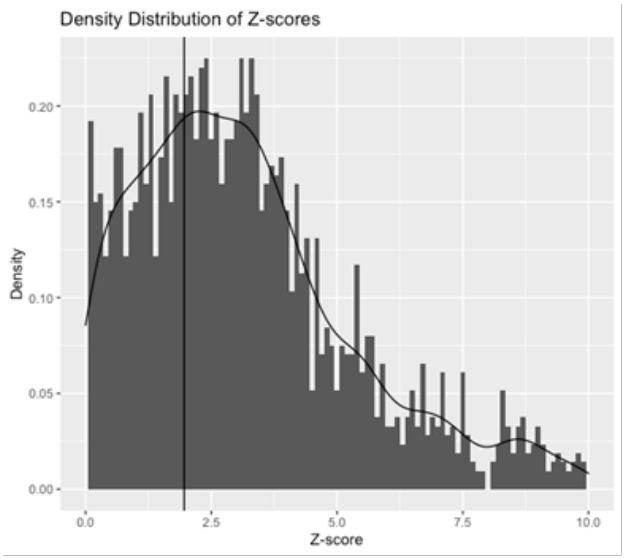

Figure 7 displays the distribution of z-scores presented in the articles replicated by Linnainmaa (2018). For full reference list see "Appendix". Bin-widths of 0.1 and a black line as a marker for a z-score of 1.96 (significance at the 0.05 level) as well as z-score of 2.2 (maximum).

The distribution of the z-scores is presented in as a histogram in figure 4. We see a presence of a maximum at roughly 2.2, which represents a p-value of under 0.05. Another observation is the denser distribution on the right hand side of the z = 1.96 level than the left hand side. There is somewhat of a steep drop off of the density line on the left hand side, as opposed to a nearly flat line up until z = 3 . After that we see a natural decrease in tests before leveling off.

From our plotted distribution, we see that the density line shares some similar characteristics to those exhibited by the plots displayed by Brodeur et al. (2016). They report a distribution with a camel-back shaped hump, with a maximum slightly above 2, but also a local minimum at z = 1.5 (p-value of 0.12). By the nature of tests, the distribution would decrease for higher z-values. This fact brings interest to our maximum at 2.2, which is similar to the maximum reported by Brodeur et al (2016); however, this could also be a result of selective reporting of the significant tests. We fail to see the clear camel-back shaped hump, where there is a dip in the distribution right before the level 1.96, that Brodeur et al. find and attribute to p-

hacking. However, we find a slight indentation at around the same location as Brodeur et al.(2016) find their local minimum. This discrepancy between their dip and our indentation could be explained in the much fewer number of articles, and therewith tests, that we used in our distribution (our 22 compared to their 642). All in all, it is hard to make the claim that the original articles were subject to p-hacking only through the inferences from our histogram. However, pieced together with the results in our replication study, it gives ground to believe that this could be the case.

6.2 Multiple-testing adjustment as an alternative to out-of-sample validation

Harvey et al. suggests the usage of a test statistic cutoff of 3.00 instead of the conventional 1.96 cutoff for a two-sided test at 5% significance. An unadjusted t statistic of 3.00 on a two-sided hypothesis test would roughly equate a p-value of 0.002. This can be compared to the 0.005 p-value proposed by Benjamin et al. (2017), as a better threshold for statistical significance. Anomalies which clear the 3.00 hurdle are regarded as significant despite the presence of multiple testing in anomalies literature, and this statistically motivated cutoff also serves as a mitigation against this form of p-hacking. Thus, to verify the robustness of an anomaly against p-hacking, implementing a multiple-testing t-statistics cutoff is an alternative to out-of-sample testing.

We find that only 6 out of 21 anomalies have in-sample average returns t-statistics that clear the 3.00 hurdle. However, to investigate multiple-testing adjustments as an alternative to out-of-sample validation we are also interested in whether the anomalies that are significant at 3.00 are the same anomalies that proved to be significant out-of-sample. Theoretically, clearing Harvey et al. (2016)'s multiple-testing cutoff should imply that an anomaly is regarded as significant despite the presence of data-snooping, and therefore that anomaly should not only be an in-sample phenomenon. To examine whether out-of-sample significance coincides with clearing the 3.0 hurdle in-sample, we present a comparison in Table 5. The table compares the results of out-of-sample validation and implementing a multiple testing t-statistic cutoff. One column specifies whether an anomaly is significant in-sample at the 3.00 cutoff, and the other column specify whether the same anomaly is significant in any of the out-of-sample periods (pre-sample, pre-publication, post-sample) at the conventional cutoff of 1.96.

Out-of-sample significance vs. statistical adjustments

Anomaly Factor	In-sample t-value > 3	Out-of-sample significance	Intersection
Abnormal capital investment	Yes	Yes	X
Accruals	No	No	X
Asset Growth	Yes	No	
Book-to-market ratio	No	No	X
Cash flow-to-price ratio	Yes	No	
Change in asset turnover	No	Yes	
Enterprise Multiple	No	No	X
Gross Profitability	No	No	X
Growth in Inventory	No	Yes	
Growth in Sales minus Inventory	No	Yes	
Investment growth rate	Yes	No	
Investment-to-assets ratio	Yes	No	
Investment-to-capital ratio	No	No	X
Net Operating Assets	Yes	Yes	X
Net Working Capital Changes	Yes	Yes	X
Operating Profitability	No	No	X
Profit margin	No	No	X
ROA	No	No	X
ROE	No	No	X
Sales-to-price ratio	No	Yes	
Sustainable Growth	No	No	X

Table 5. "In-sample t-value > 3" shows which anomalies clear the 3.00 test statistic cutoff for average returns, where "Yes" indicates that an anomaly has average return t-statistic of more than 3.00. "Out-of-sample significance" denotes significance in any of the out-of-sample periods (Pre-sample, Pre-publication and Post-sample) and "Yes" indicates that an anomaly has a t-statistic above the conventional 1.96 level in any of the three out-of-sample periods. In "Intersection", anomalies for which both out-of-sample and multiple-testing cutoff yield the same conclusion are marked with "X".

We find that only 13 out of 21 anomalies coincide with respect to the two different validation methods. Therefore, a considerable amount of anomalies are "robust" with regards to one approach but not the other. Concludingly, the multiple testing adjustment is not a substitute for out-of-sample tests. Additionally, more conclusions can be derived from out-of-sample tests due to the time-period analysis, which cannot be conducted through multiple-testing adjustments.

6.3 Analyzing anomaly performance

In our results, we pointed out that different anomaly types have different characteristics with regards to average returns and alphas. Here, we provide plausible explanations for the observed results. Profitable firms tend to be larger, less volatile and more liquid, which also implies that profitability anomalies which go long on these firms will have returns of lower magnitude while being more stable. This is an explanation for the low average return, but still significant CAPM and FF3 alpha of profitability anomalies. Valuation based anomalies seem to produce lower FF3 alpha than other anomalies overall. This can be explained by the fact that value risk premium is accounted for by the three factor model. Thus, the return predictability in valuation based anomalies are already explained by the book-to-market factor in FF3. This also explains why the book-to-market anomaly itself has negative alpha throughout all sample periods. Furthermore, some anomalies seem to be more persistent than others, which can be attributed to limits to arbitrage such as trading frictions.
Anomalies with lower arbitrage costs will suffer from larger performance decay. Chu et al (2020) show that anomalies such as gross profitability, asset growth, investments to assets, return on assets, net operating assets and accruals are largely driven by mispricing due to limits to arbitrage.

7. Concluding remarks

7.1 Conclusion

Our out-of-sample tests, z-score distribution and also tests based on multiple testing adjustments all point to that part of the existing anomalies literature can be attributed to p-hacking. Compared to Mclean & Pontiff (2015), we have a higher upper bound estimate for p-hacking effects, and a smaller effect from publication informed trading. Also, we find that only 6 of 21 anomalies clear the multiple testing t-statistic cutoff of 3.0 suggested by Harvey et al. Thus, we conclude that a majority of published accounting-based anomalies are probably false discoveries. Our study supports the idea that p-hacking and publication bias in the field of cross-section of stock returns is a major issue. Papers that support the idea that p-hacking and publication bias is dominant include Harvey et al. (2016) and Hou et al. (2017), and Linnainmaa & Roberts (2018), while papers like Mclean & Pontiff (2015) find that this is a relatively minor issue.

It is worth addressing that we do not assign the entire anomaly "factor zoo" to p-hacking either. To shed light on the limitations of p-hacking as an explanation, Chen (2019) conduct a thought experiment and argues that it would take 15 million years to find the 316 factors in the Harvey, Liu & Zhu (2016) through purely p-hacking. To conclude, we attribute part of the cross-sectional return predictability in accounting-based stock anomalies to p-hacking.

7.2 Directions for Future Research

It is valuable to study a larger set of anomalies to enhance the statistical power of the conclusions drawn from this study. Also, further investigation of the differences in computation methods between the original studies and our study would be interesting to pinpoint why some anomalies fail to replicate. It would be valuable to study anomalies through other analyses that investigate p-hacking or publication bias. An example of this would be using the method brought forth by Andrews & Kasy(2019) which estimates publication bias through meta-studies.

References

Andrews, I., & Kasy, M. (2019). Identification of and correction for publication bias. *American Economic Review*, *109*(8), 2766-94.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., ... & Johnson, V. E. (2018). Redefine statistical significance. *Nature human behaviour*, *2*(1), 6-10.

Brodeur, A., Lé, M., Sangnier, M., & Zylberberg, Y. (2016). Star wars: The empirics strike back. *American Economic Journal: Applied Economics*, *8*(1), 1-32.

Chen, A. Y. (2019). The limits of p-hacking: A thought experiment.

Fama, E. F., & French, K. R. (1992). *The cross-section of expected stock returns* (pp. 349-391). University of Chicago Press.

Fama, E. F., & French, K. R. (2008). Dissecting anomalies. *The Journal of Finance*, *63*(4), 1653-1678.

Fanelli, D. (2013). Why growing retractions are (mostly) a good sign. *PLoS Med*, *10*(12), e1001563.

French, K. R. (2008). Presidential address: The cost of active investing. *The Journal of Finance*, *63*(4), 1537-1573.

Gelman, A., & Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. *Department of Statistics, Columbia University*, *348*.

Harvey, C. R. (2017). Presidential address: The scientific outlook in financial economics. *The Journal of Finance*, *72*(4), 1399-1440.

Campbell R. Harvey, Yan Liu, Heqing Zhu, ... and the Cross-Section of Expected Returns, *The Review of Financial Studies*, Volume 29, Issue 1, January 2016, Pages 5–68

Hasbrouck, J. (2009). Trading costs and returns for US equities: Estimating effective costs from daily data. *The Journal of Finance*, *64*(3), 1445-1477.

Kewei Hou, Chen Xue, Lu Zhang, Replicating Anomalies, *The Review of Financial Studies*, Volume 33, Issue 5, May 2020, Pages 2019–2133

Linnainmaa, J. T., & Roberts, M. R. (2018). The history of the cross-section of stock returns. *The Review of Financial Studies*, *31*(7), 2606-2649.

McLean, R. D., & Pontiff, J. (2016). Does academic research destroy stock return predictability?. *The Journal of Finance*, *71*(1), 5-32.

Szucs, D., & Ioannidis, J. (2017). When null hypothesis significance testing is unsuitable for research: a reassessment. *Frontiers in human neuroscience*, *11*, 390.

Shleifer, A., & Vishny, R. W. (1997). The limits of arbitrage. *The Journal of finance*, *52*(1), 35-55.

Appendix

Anomalies
This section of the appendix will define the anomalies brought forth in section 3.2. The order of the anomalies corresponds to that in table 1. We also list the study that first used the variable to explain the cross-section of stock returns and the year the study was published.

Growth and Investment
Abnormal capital investment is defined as capital expenditures scaled by revenues, scaled by the average of this ratio over the previous three years
Titman et al. (2004) measure the predictive power of abnormal capital investment using return data from July 1973 through June 1996.

Asset growth is defined as the percentage change in total assets
Cooper et al. (2008) examine the predictive power of asset growth using return data from July 1968 to June 2003.

Growth in inventory is defined as the change in inventory divided by the average total assets
Thomas and Zhang (2002) use growth in inventory to predict stock returns using return data from 1970 to 1997.

Growth in sales minus inventory is the difference between sales growth and inventory growth. Sales growth is the increase in sales over its average value over the previous two years, all scaled by the average value over the previous two years; inventory growth is the increase in inventory over its average value over the previous two years, all scaled by the average value over the previous two years
Abarbanell and Bushee (1998) use growth in sales minus inventory to predict returns from 1974 through 1993.

Investment growth rate is the percentage change in capital expenditures
Xing (2008) uses investment growth rate to construct an investment factor using return data from 1964 to 2003.

Investment-to-assets ratio is defined as the change in the net value of plant, property, and equipment plus the change in inventory, all scaled by lagged total assets
Lyandres et al. (2008) use the investment-to-assets ratio to predict returns from January 1970 through December 2005

Investment-to-capital ratio is defined as the ratio of capital expenditures to the lagged net value of plant, property, and equipmen

Xing (2008) uses the investment-to-capital ratio to predict stock returns using return data from 1964 to 2003.

Sustainable growth is defined as the percentage change in the book value of equity
Lockwood and Prombutr (2010) use return data from July 1964 through June 2007 to measure the predictive power of sustainable growth.

Earnings Quality

Accruals is the noncash component of earnings divided by the average total assets
Sloan (1996) uses data from 1962 to 1991 to examine the predictive power of accruals.

Net operating assets represent the cumulative difference between operating income and free cash flow, scaled by lagged total assets
Hirshleifer et al. (2004) form trading strategies based on net operating assets using data from July 1964 through December 2002.

Net working capital changes is another measure of accruals
Soliman (2008) uses net working capital changes to predict stock returns using return data from 1984 to 2002

Profitability

Change in asset turnover is defined as the annual change in asset turnover, where asset turnover is revenue divided by total assets
Soliman (2008) uses the change in asset turnover to predict returns between 1984 and 2002

Gross profitability is defined as the revenue minus cost of goods sold, all divided by total assets
Novy-Marx (2013) examines the predictive power of gross profitability using return data from July 1963 through December 2010

Operating profitability is defined as the revenue minus cost of goods sold, SG&A, and interest, all divided by book value of equity
Fama and French (2015) construct a profitability factor based on operating profitability using return data from July 1963 through December 2013

Profit margin is defined as the earnings before interest and taxes, divided by sale
Soliman (2008) uses profit margin to predict returns using return data from 1984 to 2002.

Return on assets is defined as the earnings before extraordinary items, divided by total assets
Haugen and Baker (1996) use return on assets to predict returns between 1979 and 1993

Return on equity is defined as the earnings before extraordinary items, divided by the book value of equity
Haugen and Baker (1996) use return on equity to predict returns between 1979 and 1993

Valuation
Book-to-market ratio is defined as the book value of equity divided by the December market value of equity
Fama and French (1992) use book-to-market ratio to predict returns using return data from July 1963 through December 1990

Cash flow-to-price ratio is defined as the income before extraordinary items plus depreciation, all scaled by the December market value of equity
Lakonishok et al. (1994) use the cash flow-to-price ratio in tests that use return data from May 1968 through April 1990

Enterprise multiple is a value measure used by practitioners
Loughran and Wellman (2011) compare the predictive power of enterprise multiple to that of book-to-market using return data from July 1963 through December 2009

Sales-to-price ratio is defined as total sales divided by December market value of equity
Barbee et al. (1996) compare the predictive power of sales-to-price to those of book-to-market and debt-to-equity ratio using return data from 1979 through 1991

Z-score histogram
In this section we will list the articles from which we collected the data to create our histogram plot.

Abarbanell, J. S., & Bushee, B. J. (1998). Abnormal returns to a fundamental analysis strategy. *Accounting Review*, 19-45.

Bhandari, L. C. (1988). Debt/equity ratio and expected common stock returns: Empirical evidence. *The journal of finance*, *43*(2), 507-528.

Bradshaw, M. T., Richardson, S. A., & Sloan, R. G. (2006). The relation between corporate financing activities, analysts' forecasts and stock returns. *Journal of accounting and economics*, *42*(1-2), 53-85.

Campbell, J. Y., Hilscher, J., & Szilagyi, J. (2008). In search of distress risk. *The Journal of Finance*, *63*(6), 2899-2939.

Cooper, M. J., Gulen, H., & Schill, M. J. (2008). Asset growth and the cross-section of stock returns. the Journal of Finance, 63(4), 1609-1651.

Daniel, K., & Titman, S. (2006). Market reactions to tangible and intangible information. *The Journal of Finance*, *61*(4), 1605-1643.

Dichev, I. D. (1998). Is the risk of bankruptcy a systematic risk?. *the Journal of Finance*, *53*(3), 1131-1147.

Haugen, R. A., & Baker, N. L. (1996). Commonality in the determinants of expected stock returns. *Journal of financial economics*, *41*(3), 401-439.

Hirshleifer, D., Hou, K., Teoh, S. H., & Zhang, Y. (2004). Do investors overvalue firms with bloated balance sheets?. *Journal of Accounting and Economics*, *38*, 297-331.

Hou, K., & Robinson, D. T. (2006). Industry concentration and average stock returns. *The Journal of Finance*, *61*(4), 1927-1956.

Lakonishok, J., Shleifer, A., & Vishny, R. W. (1994). Contrarian investment, extrapolation, and risk. *The journal of finance*, *49*(5), 1541-1578.

Lockwood, L., & Prombutr, W. (2010). Sustainable growth and stock returns. *Journal of Financial Research*, *33*(4), 519-538.

Loughran, T., & Wellman, J. W. (2011). New evidence on the relation between the enterprise multiple and average stock returns. *Journal of Financial and Quantitative Analysis*, 1629-1650.

Lyandres, E., Sun, L., & Zhang, L. (2008). The new issues puzzle: Testing the investment-based explanation. *The Review of Financial Studies*, *21*(6), 2825-2855.

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

Pontiff, J., & Woodgate, A. (2008). Share issuance and cross-sectional returns. *The Journal of Finance*, *63*(2), 921-945.

Sloan, R. G. (1996). Do stock prices fully reflect information in accruals and cash flows about future earnings?. *Accounting review*, 289-315.

Soliman, M. T. (2008). The use of DuPont analysis by market participants. *The Accounting Review*, *83*(3), 823-853.

Spiess, D. K., & Affleck-Graves, J. (1999). The long-run performance of stock returns following debt offerings. *Journal of Financial Economics*, *54*(1), 45-73.

Titman, S., Wei, K. J., & Xie, F. (2004). Capital investments and stock returns. *Journal of financial and Quantitative Analysis*, *39*(4), 677-700.

Xing, Y. (2008). Interpreting the value effect through the Q-theory: An empirical investigation. *The Review of Financial Studies*, *21*(4), 1767-1795.

Figures

This section provides supplementary figures.

Figure 8: Average Return t-statistics by sample period

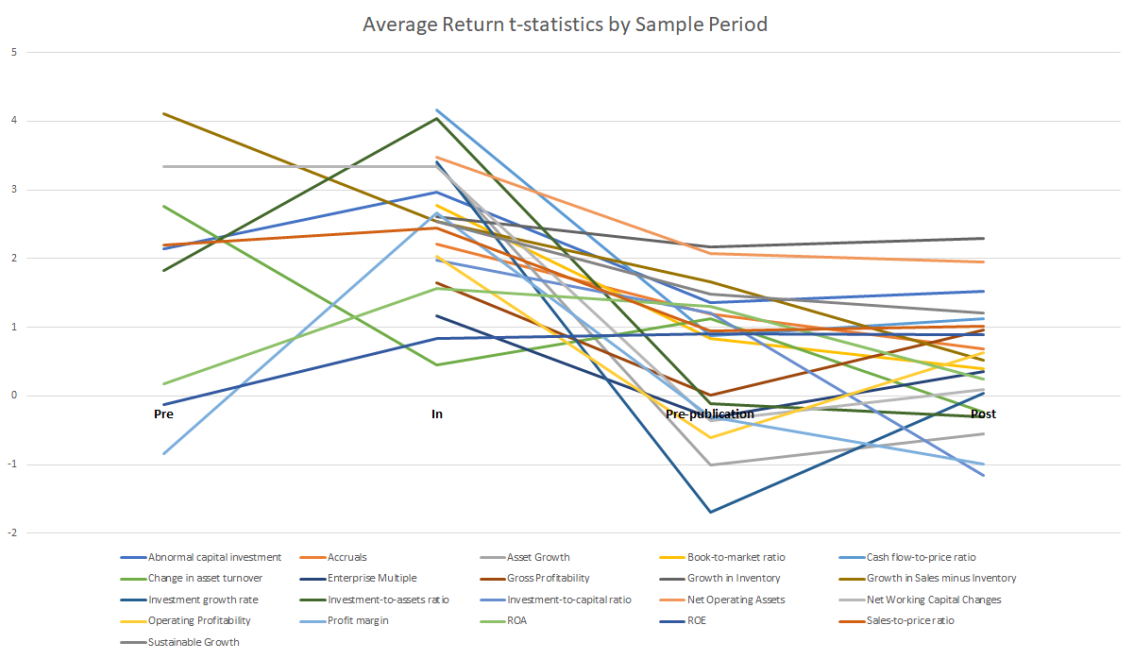

Figure 8 plots the t-statistics of average monthly returns by sample period and anomaly.

Figure 9: Capital asset pricing model alpha t-statistics by sample period

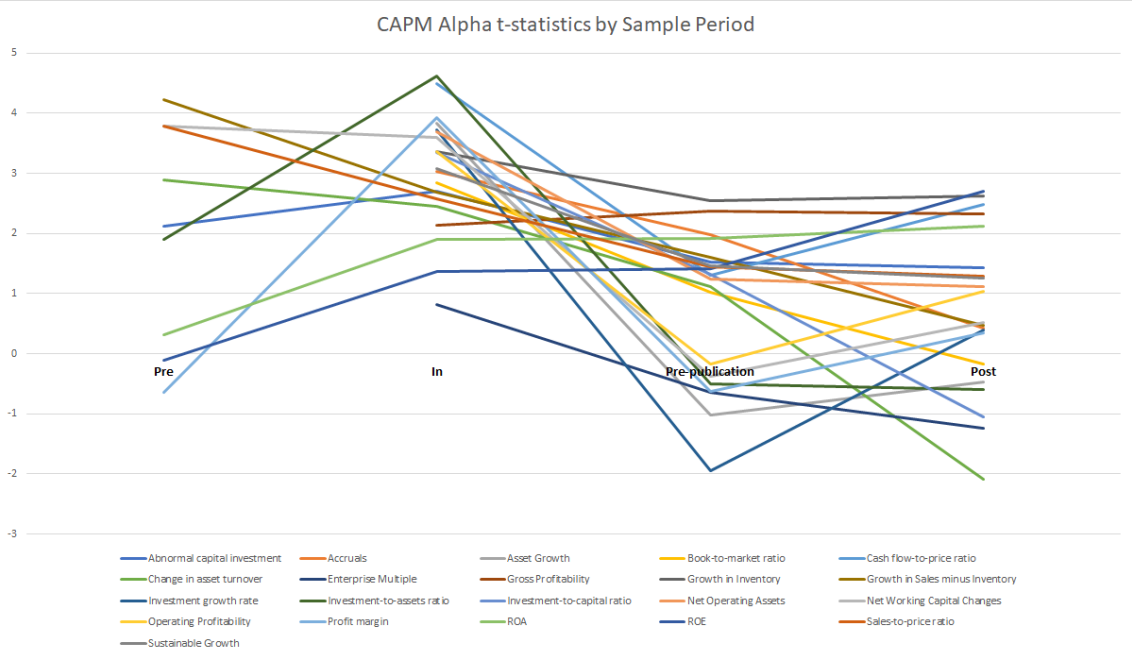

Figure 9 plots the t-statistics of CAPM Alpha by sample period and anomaly.

Figure 10: Fama-French three factor model alpha t-statistics by sample period

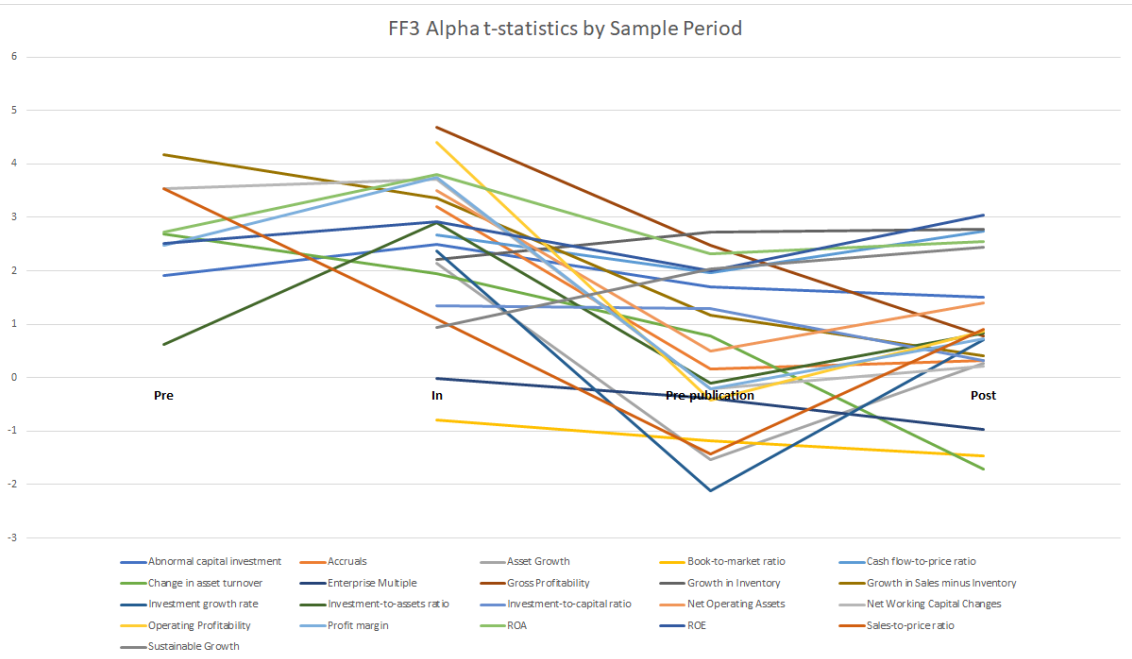

Figure 10 plots the t-statistics of FF3 Alpha by sample period and anomaly.