Essays in Empirical Asset Pricing and Behavioral Finance

Inaugural Dissertation
submitted in partial fulfillment of the requirement for the degree of
Doctor rerum oeconomicarum (Dr. rer. oec.)
at the Schumpeter School of Business and Economics
of the University of Wuppertal

submitted by
Jan Philipp Harries

Wuppertal, June 2021

The PhD thesis can be quoted as follows:

urn:nbn:de:hbz:468-urn:nbn:de:hbz:468-20211223-094524-0
[http://nbn-resolving.de/urn/resolver.pl?urn=urn%3Anbn%3Ade%3Ahbz%3A468-20211223-094524-0]

DOI: 10.25926/hc4g-8b08
[https://doi.org/10.25926/hc4g-8b08]

Preface

The first time I came into contact with financial markets was during high school, when participating in a stock exchange game. While I couldn't put my finger on it at that time, I was instantly fascinated by the way financial markets work. Before, it never came to my mind that prices of goods and assets in our economy are not a constant, set arbitrarily by some seller like a business owner or the government, but the direct product of supply and demand. And that markets of all kinds are extremely important for the functioning of the economy, as they absorb the individual knowledge of all individuals in a society to facilitate the price discovery process and thus help determining the value of a given asset.

Intrigued by that insight, it was no coincidence - at least in hindsight - that I ended up trading stocks and options in the college dorm during my undergraduate studies[1] and started working on the Equity Trading floor of a bank shortly thereafter. While still admiring the elegance of price discovery purely through supply and demand, I also realized that I didn't really understand what was going on much of the time. Additionally, in the aftermath of the financial crisis it became clear to me that markets - if left unchecked - can also produce weird and partially undesirable results.

Thus, I set out to learn more about capital markets and their efficiency. This thesis is the result of four years of research on various topics in the areas of empirical asset pricing and behavioral finance that are related to capital market efficiency in one way

[1] As this was shortly before the financial crisis of 2008, I also learned about bubbles and risk management the hard way!

or another. I hope that the papers that are part of this dissertation constitute a small but meaningful contribution to their respective areas of research.

My time as a research associate in Wuppertal was not only enlightening in terms of scientific insights but also due to the many fantastic people I met. While my name is listed on the cover, this dissertation wouldn't have been possible without their support.

First and foremost, I want to thank my supervisor André Betzer. He not only encouraged me to pursue graduate studies and agreed to supervise my dissertation but also motivated me throughout the entire project and provided invaluable support and advice. Additionally, I am especially grateful to Nils Crasselt, who not only consented to be the second supervisor but has also supported me in countless ways from the early stages of this dissertation.

I want to thank my coauthors Stephan Kessler and Bernd Scherer for their willingness to work with me on (so far) two very intense research projects, during which I learned a ton.

Iris Leclaire and Anne Peuyn helped me in many, many ways during my time in Wuppertal and always knew the right way to make things happen, for which I am very grateful. I want to thank Stefan Thiele, Paul J.J. Welfens and Peter Witt for their willingness to become members of my dissertation committee. Additionally, I want to thank all my colleagues at the chair of controlling and the chair of finance and corporate governance for the good working atmosphere and the many productive and interesting debates during our PhD seminars. I am particularly grateful to (in alphabetical order) Dmitry Bazhutov, Christian Danisch, Markus Doumet, Christian Lohmann, Steffen Möllenhoff and Sascha Schworm for their support during different stages of this dissertation.

I also want to thank André Kuck, who supervised my undergraduate thesis and who encouraged me to follow my intellectual curiosity and leave the corporate world for academia and my good friend Christoph Pott, who also consistently motivated me and managed to find more orthographic errors and little inconsistencies in my drafts than I would've imagined.

Finally, I want to thank my family. My parents always supported me, not only during the whole endeavour of pursuing a PhD, and I wouldn't be the same person without their love and advice.

But most of all, I have to thank my wife Anika, who supports me in every imaginable way and always manages to cheer me up in times of discouragement. Without her by my side, this dissertation would not have been possible. And although becoming a father during this project has probably not made it easier, Anika has shielded me from a lot of the stress in early parenthood. Now, one smile of our son Jakob is always enough to forget everything else and the greatest motivation to carry on.

Jan Philipp Harries
Düsseldorf, June 2021

Note: The second chapter, "Value by Design", is not contained in the published version of this dissertation due to copyright reasons. The article was published in the Journal of Portfoliomanagement Quantitative Special Issue 2020, 46 (2) 25-43 and can be accessed at `https://doi.org/10.3905/jpm.2019.1.122`

Contents

List of Tables

List of Figures

1 Introduction

In economic sciences, Hayek (1945) was one of the first to explicitly highlight the crucial aspect of information absorption and dispersion in his influential essay *The Use of Knowledge in Society*: "We must look at the price system as such a mechanism for communicating information if we want to understand its real function". He argued that open markets are the best economic solution for the "problem of the utilization of knowledge which is not given to anyone in its totality" and superior to any central planning in that regard.

Modern financial markets and specifically stock exchanges represent the information-absorption and price-finding process in its purest form. Publicly-traded companies and stock exchanges are no recent invention; while first predecessors of stock exchanges probably already existed during the Roman Republic, many historians see the developments in the early seventeenth century with the foundation of the Dutch East India Company and a few years later the Amsterdam Stock Exchange in 1609 (see e.g. Baskin, 1988) as the origin of today's stock exchanges. Since then, many economists explored the inner workings of these markets in great detail. Modern data collection techniques and statistics allowed scientists in more recent times unprecedented insights into the price discovery process at stock exchanges. Bachelier (1900) is most often credited with setting the agenda for research on how prices on stock exchange behave. He stated that "past, present or even anticipated" events are already reflected in stock prices and was the first to describe the remaining price fluctuation in terms of a random-walk model, using a Brownian motion process. His work gained popularity in the 1960s and different researchers expanded upon Bacheliers ideas (amongst others, see e.g. Mandelbrot, 1963; Fama, 1965; Samuelson, 1965) until Fama (1970), who specified, formalized and empirically tested the Efficient-market hypothesis (EMH), which is commonly seen as the foundation of risk-based asset pricing and a cornerstone of capital market research. In a nutshell, the theory states that all available information is reflected in asset prices and thus no investor can consistently achieve above-average risk-adjusted returns in absence of new information.

While the validity of the EMH was (and still is) heavily debated, its development led to a breakthrough in the research of asset prices. The capital asset pricing model (CAPM), independently developed by Sharpe (1964), Lintner (1965) and Black (1972), divides risk into systematic (non-diversifiable) and unsystematic portions, which allows investors to compare risk-adjusted returns and enables modern portfolio theory. Shortcomings of the CAPM in turn led to the development of Fama and French's (1993) three-factor model which adds factors on size and valuation for a better explanation of portfolio returns.

In one way or another, all three papers in this dissertation are contributions exploring different aspects of the efficiency of capital markets. The first two papers, *Value by Design?* and *The Choice* are closely related to the three-factor model by Fama and French (1993). In *Value by Design?* we focus on the value factor (in the literature called *HML* for "high-minus-low" book-to-price portfolios) of that model, which has spawned its own investing style. With a focus on applications in investing and asset management, we show that, even though Fama and French's (1993) model uses the book-to-price ratio for its HML factor, there is still no consensus on what *value investing* exactly encompasses. Depending on different design choices, returns of different value strategies vary considerably. We also identify a *hierarchy of choices* which will be useful in investment strategy development. Additionally, we contribute to research governance, extending methods to assess the degrees of freedom consumed in backtesting and derive adjusted t-values that prevent overfitting.

The Choice builds on this idea and shifts the focus from asset management to asset pricing research. Concentrating on seemingly innocent choices around universe selection, breakpoint selection, rebalancing frequency, weighting decisions and others, we evaluate how these minor differences in factor design influence the performance and explanatory power of Fama and French's (1993) three-factor model. Replicating and extending their results, we also investigate how results changed post-publication and whether design choices make a difference in out-of-sample performance. Additionally, we analyze the results from an investor's point of view.

The third paper, *If he's still in, I'm still in!* touches a slightly different aspect in connection with the efficiency of capital markets. Instead of Asset Pricing, this paper takes a behavioral finance perspective. While markets work perfectly under the assumptions of complete and always-available information and rational market participants, these preconditions are rarely present in the real word. Bondt and Thaler (1985) found that "most people tend to “overreact” to unexpected and dramatic news events" and provided empirical evidence of substantial weak-form market inefficiencies caused by the psychology of individual decision making. Following them, scholars found many examples of behaviors that are inconsistent with the rationality assumed by the EMH. In our paper, we add a very recent angle to this literature, using the surge in Gamestop shares in early 2021 as an example to show how Social Media posts affect retail trading. This situation poses also as an example for situations where markets can fail without proper boundaries or regulation.

The fourth and final paper, *Determinants of Blockholdership*, puts two very important issues of (economic) research in the foreground: reproducibility and data availability. While planning to analyze the effects of long-term ownership on firm performance using an Asset Pricing approach, I discovered that no publicly available and reliable dataset for blockholder data in the US (specifically Form 13D and Form 13G SEC filings) exists. By writing a parser and open sourcing the dataset, I hope to stimulate further research in this area. Additionally I use the data to analyze determinants of blockholdership from a valuation perspective.

In the following paragraphs, I will briefly summarize all four papers.

1.1 Value by Design?

Value by Design was written in September 2019 together with my coauthors Dr. Stephan Kessler and Dr. Bernd Scherer. Besides their academic activities, they both work in quantitative asset management and noted that although most definitions of *value* can ultimately be traced back to Fama and French's (1993) HML factor, there is no real consensus in industry what *value* actually means. Today, there is a myriad of different *value* investment funds and factors available to invest in, which differ in many ways. Thus, our paper has the goal to further our understanding of value investing. We provide a new angle by describing value investing as the union of possible implementations of commonly-used value metrics and portfolio construction approaches, providing an envelope for value investing instead of relying on one specific definition. The paper is mainly targeted at an audience interested in implementing an equity value strategy and thus mainly focuses on strategy risk-return characteristics instead of pricing errors.

To dissect value as an investment style, we demonstrate some of the choices that investors and researchers face when creating long-short value portfolios. Our first choice is the value signal definition to be used. With the book-to-price ratio, used by Fama and French (1993) as a starting point, common alternatives are dividend yield (e.g. Lakonishok et al., 1994), price-to-earning ratios (e.g. Basu, 1983) or cash flow-related metrics (e.g. Chan et al., 1991). The second choice relates to the weighting of individual stocks. Fama and French (1993) ranked stocks for the signal and then sorted them into equally-weighted portfolios by defining percentile-based (30%) breakpoints. Other possible choices are e.g. different percentiles or more complex weighting schemes, for example using logistic or logit weightings or the ranking-based linear weighting approach by Bender and Wang (2016). Our third choice centers around the implementation of short positions. Many investors short a market index rather than individual equities, as Fama and French (1993) did for their long-short portfolios. Another choice is the handling of residual market risks or industry risks, which poses the question of whether net industry exposure reflects additional noise or has to be regarded as an important part of the strategy due to sectoral valuation effects. Our final choice is the portfolio rebalancing frequency which can range from monthly to annual and addresses the speed of information decay. The variations and their calculation are described in more detail in the chapter *Summary of Investment Strategy Dimensions and Related Choices* and appendix A.

These choices then result in a set of 3,168 strategy risk-return profiles that allow us to get a better understanding of strategy dispersion and to compute better confidence bands around Sharpe ratio estimates. Many researchers argue about whether to explain the origins of a value premium as violation of the EMH or reward for systematic risks (e.g. Barberis and Thaler, 2003; Shleifer and Vishny, 1997; Shleifer, 2000), but most found significantly positive returns for a systematic value strategy. We confirm these results, finding that even before taking into account investor skill[1], almost all permutations deliver excess returns with an average Sharpe-ratio of 0.33. A more detailed performance review can be found in the chapter *Impact of Investment Strategy Design Choice on Performance.*

An additional contribution of our paper is a "hierarchy of choices" for value investing. Our results show that cash flow- and earnings-related metrics provide the best and the

[1] Which we measure as the investor's ability to choose the right parameters for the implementation of value investing.

dividend yield metric provides the lowest risk-adjusted returns. Returns appear to be mainly driven by the long portfolios which mirrors the findings by Ammann et al. (2011) for US equity momentum strategies. Neutralizing sector exposure turns out to have a positive effect on strategy performance while different signal weighting schemes uniformly lead to worse results. The main results can be found in the table *Impact of Design Choices*.

Finally, we replace conventional, purely statistically-motivated tests for data snooping with an approach, that takes the potential data snooping bias embedded in the design decisions into account. Following Harvey and Liu (2015), we confirm that the degrees of freedom embedded in strategy design require investors to adjust their t-statistic thresholds substantially. For our choices, investors should require a t-statistic of 3.72 to arrive at the 2.5% confidence level, which means that strategies with a Sharpe ratio of below 0.69 fail to reach statistical significance. For researchers and investors, we provide an easy-to-follow procedure to haircut strategy results and account for overfitting in strategy research in the chapter *Design Choice and Significance Testing*. In the final chapter, *Design Fishing*, we also demonstrate that design fishing, e.g. the selection of choices purely by statistical analysis, can deliver significantly positive out-of-sample returns. Selecting strategies with an SSPA test proposed by Hsu and Hsu (2006) as well as naively by sharpe ratios results in out-of-sample sharpe ratios of up to 1.59.

While *Value by design* solely focused on value investing and value risk premiums, the results of this paper motivated us to also analyze other risk premiums in a similar way and to not only focus on individual factors, but also on their interaction. Thus, in our follow-up paper *The Choice*, we broadened the perspective to include asset pricing effects instead of focusing on investment-specific consequences only and analyzed, how asset design uncertainty and parameter choices affect the asset pricing performance of Fama and French's (1993) three-factor model.

"Value by Design" has been published in the Quantitative Special Issue 2020 of the Journal of Portfolio Management. The paper has been awarded the 2020 EQDerivatives Award for the "Systematic Investing Research Paper of The Year" and the results were covered by Bloomberg[2] *and in the Institutional Investor*[3].

1.2 The Choice - Reviewing the Impact of Modeling Choices on the Fama and French Three-Factor-Model

Fama and French's (1993) three-factor model is commonly seen as the foundation of modern asset pricing. After Fama and French (1992) elaborated on some weaknesses and inconsistencies of the CAPM, they were among the first who empirically attributed cross-sectional stock returns to multiple, economically-motivated tradeable and priced risk factors. Today, almost 30 years after its publication, their model is still widely used by scholars and practitioners. Since then, many studies examined how well the model really performs by analyzing pricing errors for different universes and periods. Discovered weaknesses of the model and new insights led to the development of many alternative models and additional factors claiming to improve upon the three-factor model and its explanatory power for the cross-section of stock returns. Recent examples in the literature include, for example, Hou et al. (2019), Kan et al. (2019), Hanauer (2020) or even Fama and French (2015) themselves with their own expanded five-factor model. Instead of adding to this path of the literature, we take a second

[2] See `https://www.bloomberg.com/news/articles/2020-01-16/quants-show-they-re-still-human-with-3-168-versions-of-value`.

[3] See `https://www.institutionalinvestor.com/article/b1p62z599ns4pd/The-Sharpe-Ratio-Broke-Investors-Brains`.

look at the inner workings of the original three-factor model, focusing on neglected research questions around factor design and subtle design choices.

While the performance of the three-factor model and others has been thoroughly analyzed in the literature(compare e.g. Harvey et al., 2016; Fama and French, 2018; Hou et al., 2020), we found that there was significantly less attention paid to the small and seemingly unimportant choices made by Fama and French and other researchers during the factor calculation. As we had seen in our previous paper, *Value by Design?*, these choices can make a huge difference in results, even when the general model and underlying assumptions stay the same.

To study the effect of researcher choices, we build thousands of different three-factor models around nine variations:

- The merge date (which describes when a company's book value is assumed to be available),
- the marketcap date (which determines whether the dates for stock price data and fundamental data are synchronized for the calculation),
- the universe used for calculation of breakpoints, as well as long-short portfolios,
- the percentiles or breakpoints used for calculation of the different HML and SMB portfolios,
- the rebalancing frequency,
- the use of winsorization and
- the weighting scheme.

The motivation for these variations as well as a review of existing literature on these can be found in chapter 3.2 and a detailed explanation of all variations with their possible values can be found in chapter 3.3, with figure 3.1 depicting our factor-creation process.

Our results confirm a large dispersion in the risk-return characteristics of individual factors for alternative implementation choices, with in-sample Sharpe ratios ranging from 0.22 to 1.01 for HML (vs 0.52 for the original three-factor model) and from -0.13 to 0.46 for SMB (vs 0.32 for the original three-factor model). Hereby, the choices which most significantly affect results are the universe for breakpoint-calculation as well as portfolios, the use of winsorization and the weighting method. Looking at all three factors, we find that when adding Size and Value factors jointly to the market portfolio, they turn out to be economically meaningful, yielding higher risk-adjusted returns and lower pricing errors independent from design choices. However, this result comes with limited statistical significance after our bootstrapping procedure to adjust for the effect of data mining. From an asset pricing perspective, we find that in-sample *the choice*[4] does well with a lower pricing error than 84% of alternative implementations.

Strikingly, we find a strong negative relationship between the performance of HML and SMB across different design choices (see figure 3.2), which hasn't been discussed in the literature before. This means that, compared to single factors, maximum Sharpe ratios for complete factor models are less sensitive to design choices. It follows that design choices not only affect the relative performance of different factors in a model, but also induce a comparison issue when testing newly proposed factors and anomalies in combination with existing factor models that don't adhere to the same choices. More detailed results can be found in chapter 3.4 for individual factors and in chapter 3.6 for complete factor models. Chapter 3.5 adds an investment perspective, complementing the results of our previous paper.

Besides curiosity about the effect of these choices on model results, our study was also motivated partially by the "replication crisis" in economics (e.g. Ioannidis et al., 2017). At least since Gelman and Loken's (2013) *garden of forking paths*, researchers are increasingly aware of multiple comparison problems and that results or often sensitive

[4] Throughout the paper, we call the variations that resemble Fama and French's (1993) version of the FF3F *the choice*.

to small, conscious or unconscious choices in research design. Huntington-Klein et al. (2021) for example found that scientists are often not able to replicate empirical results in applied microeconomics even when given identical data and instructions, due to not-documented researcher choices. In Asset Pricing, for example Asness and Frazzini (2013) or Harvey (2017) warned about these problems and demonstrated the effect of some of these choices. To examine whether the three-factor model suffers from similar problems, we compare the choices Fama and French could have taken using only information that was available at publication of *the choice* and evaluate them by calculating out-of-sample results. Hereby, we are mainly interested in the persistence of in-sample design fishing while the well-documented post publication decay in factor returns (compare e.g. Mclean and Pontiff, 2016) is less interesting for us. Would Fama and French publish a new three-factor model today using only data from the time after their initial publication, would they arrive at the same design choices?

We find that Fama and French's (1993) *choice*, when compared with other design choices, does not lead to significantly better (or worse) risk-adjusted returns out-of-sample. Additionally, our results indicate that outperformance, measured by the Sharpe ratio, was apparently not a main driver of the design choices made by Fama and French. Unsurprisingly, the original specification seems to be derived from the perspective of asset pricing research, minimizing pricing errors for a cross-section of test portfolios. While its in-sample Sharpe ratio is about average, it outperforms 72.3% of alternative design choices when measuring pricing error. Details can be found in table 3.6. Out-of-sample however, the ability of all choices to price the cross-section of test portfolios is decaying equally and Fama and French's (1993) *choice* does not exhibit any particular bias. While we find some performance persistence for model performance when measuring Sharpe ratios (confirming the statement by Harvey et al., 2016, that "the optimal amount of data mining is not zero"; results can be found in table 3.5), we can't find any persistence for pricing errors, showing that missing factors can not be compensated with specific design choices.

Finally, we were surprised to find that only 40% of variations offer higher out-of-sample risk-adjusted returns than an investment into the market portfolio. Albeit this evidence comes with low statistical significance, it warrants further attention, especially in the light of the disappointing performance of many factor models in recent times.

All in all, this paper delivers new insights about the sensitivity of factor models, especially the three-factor model, to small and often overlooked design choices. While these choices can have a significant impact on results, we don't find any evidence for overfitting or bias by Fama and French (1993) and conclude that their models popularity is ex-post supported by clever design choices. However, good choices can't prevent performance decay out-of-sample.

"The Choice" has been submitted for publication in a reputable economic journal.

1.3 If he's still in, I'm still in! - How Reddit posts affect GameStop retail trading

When the stock price of GameStop Corporation increased by more than 3000% in January 2021, this violated much of what is supposed to be known about how stock markets work. There was no news that was dissipated by the market and could've caused this increase in the stock price. No earnings release, no takeover rumors and no management change was announced. Instead, financial media and regulators attributed the largest part of that enormous move of the stock price to users of r/WallStreetBets, a forum on social media site Reddit[5]. Even the chair of the SEC, Gary Gensler, stated in May 2021 that "this winter's events also highlighted the rapidly changing face of

[5] Found at `https://www.reddit.com/r/wallstreetbets/`.

social media and its intersection with our capital markets". The goal of this paper is to empirically establish a relationship between social media posts on Reddit and retail trading activity.

In 2020, users of the Reddit community r/WallStreetBets discovered that an unusual large amount of GameStop shares (above 100% of outstanding shares) was sold short by several hedge funds and pushed each other to buy as much shares and options of GameStop as possible, resulting in a so-called gamma- and short squeeze. While "Social Trading" isn't a new phenomenon, the magnitude of movements in GameStop's share price after it gained attention on Reddit is unprecedented.

Similar to other social media platforms, users on Reddit are able to post content which in turn can be commented on by other users. At the end of the first quarter of 2021, r/WallStreetBets is one of Reddit's largest communities with more than 10 million subscribers. It was created almost ten years ago and focuses explicitly on speculative equity trading. As the community emphasizes speculative trading and ”gambling”, we assume that retail trading activity by users of r/WallStreetBets may be different in nature compared to average retail equity investors. However, since retail trading on the US equity market has enormously grown since 2020, mainly due to easily-accessible broker apps like RobinHood, this fact doesn't harm the contribution of our results. Bradley et al. (2021) already showed that specific posts on r/WallStreetBets can lead to significant abnormal returns. We extend their approach to the GameStop situation by using a larger and more current dataset of all posts and considering multiple measures of retail trading activity. More details and previous literature on Reddit and r/WallStreetBets can be found in chapter 4.2.3.

Due to more easily available datasets and computational progress, researchers were able to analyze the trading of retail investors on the stock market in great detail over the last 15 years. Main focus of the literature in this area are the characteristics of retail trades as well as the predictability of retail investor trading (see e.g. Kaniel et al., 2008; Han and Kumar, 2013; Boehmer et al., 2021). Researchers also established different measures for the share of retail trading, also called retail trading proportion. In our paper, we use measures that are based on odd-lots, trade size and subpenny improvements and additionally introduce a new measure based on small-size option flows. While Han and Kumar (2013) and others already showed that stocks with a high share of retail traders often exhibit lottery-like features and that retail investors with strong gambling propensity are drawn to high-volatility stocks, we are, to the best of our knowledge, the first to empirically show how social media activity affects the retail trading proportion over time. The situation around GameStop in early 2021 is uniquely suited to establish and analyze this relationship and additionally compare different measures for retail trading activity in a high-volatility environment. We review the literature on retail trading in chapter 4.2.1 and introduce the different proxies for the retail trading proportion used in our analysis in chapter 4.3.2.

Besides retail trading, the role of short- and long-term sentiment in financial markets has been thoroughly studied before and deviations from rational investment decisions and the EMH are well-known and generally accepted[6]. Daniel et al. (2002), for example, found that psychological biases affect investor behavior and prices. Following them, other researchers demonstrated that newspaper sentiment can lead to pricing anomalies (e.g. Tetlock, 2007; García, 2013; Kumar et al., 2020) or analyzed the influence of Twitter posts on the stock market (e.g. Behrendt and Schmidt, 2018; Broadstock and Zhang, 2019; Nisar and Yeung, 2018). More Background on the effect of sentiment on the stock market can be found in chapter 4.2.2.

Putting everything together, we use a self-scraped dataset of more than 40 million Reddit posts which is then merged with tick-level stock and option price data to show that Reddit posts lead to increased retail trading activity in GameStop shares. Our results show that an increase of 50% in Reddit comments on GameStop will lead to an increase of the retail trading proportion of approximately 0.7% for shares and 0.6%

[6] The extent and duration of these deviations although are debated feverishly in the literature.

for options in the following 30-minute window. While economically small, the effect is larger than what was found in comparable studies and robust and consistent over multiple retail trading classifications. The direction of this effect was confirmed by the results of a Granger causality test. Preliminary results indicate that the effect seems to be even stronger in times of very high volatility but further subgroup analysis was out of the scope of this paper. All results can be found in chapter 4.4.

Additionally, we compare multiple procedures to separate retail from institutional trades in this timely real-world case study. Besides the aforementioned measures based on odd-lots, trade size and subpenny improvements, we introduce a novel option-based measure relying on one-contract trades that hasn't been used in this context before. While all our measures for retail trading exhibit a high correlation and seem to be well-suited to identify retail trading volume, our results indicate that more conservative measures like the one based on marketable orders introduced by Boehmer et al. (2021) could partially fail to identify retail flow in high-volatility situations like our case study. The new option-based measure for retail trading on the other hand seems to capture "Reddit-like" retail flow better than these traditional stock-based measures as evidenced by highly-significant and robust results in our regression analysis.

Further cross-sectional analysis and a cross validation of this measure with individual-level broker data would be necessary to confirm whether this holds true outside of this case study as well. However, we hope that our results on social media-induced trading and the measurement of retail trading proportion with option flows will help academics, regulators and investors to track and analyze similar developments in the future.

"If he's still in, I'm still in!" was published as a working paper in May 2021 and has been submitted for publication in a reputable economic journal.

1.4 Determinants of Blockholdership - A new Dataset for Blockholder Analysis

In the US, every exchange-listed company and many market participants have to submit specific regulatory filings to the SEC, that will then publish these in a freely-available online database called EDGAR[7]. Many of these filings have to adapt a machine-readable, standardized format which is one of the main reasons for the good data availability of fundamental data for US-listed firms or holdings data for US-centric financial investors. However, while working on a planned research project to analyze the effects of long-term stock ownership on firm performance using an Asset Pricing approach, I noticed that several filings are not required to be filed in a unified and machine-readable format, for example the Form 13D and Form 13G blockholder filings. These forms need to be submitted by every investor (in contrast to the well-known Form 13F filings which only need to be filed by financial investors above a certain size) when his stake in a publicly listed company exceeds 5%. While a rough template for these kind of filings exists, most filings differ significantly in their structure and wording. This is also the main reason for the lacking availability of the data on non-financial blockholders for researchers, which has also been adressed before in the literature (see e.g. Dlugosz et al., 2006).

After discovering that there is no suitable data source for research on original blockholder filings, I built a software to download and parse these filings myself. With this paper, I introduce and release the resulting new dataset for information contained in Form 13D and Form 13G filings. After manually scanning hundreds of different filing formats, I developed a parser that is sufficiently accurate and robust to parse the important information out of almost all blockholder filings (some details of the inner workings are described in chapter 5.4.1). As of June 2021, there are 758,666 blockholder filings from November 1993 to May 2021 contained in this database, each with

[7] Found at `https://www.sec.gov/edgar/searchedgar/companysearch.html`.

76 fields containing various information, from the reported ownership percentage to the addresses of the filing entity and the subject company. Descriptive statistics and an overview of the data is given in chapter 5.4. I hope that the free availability of this data enables other researchers to get a better understanding of some important topics in financial markets. For topics like e.g. the ongoing discussion about the advantages of long-term oriented shareholders, which is not restricted to academic circles[8], data availability is crucial to enable and support empirical conclusions.

As a second contribution of this paper, I conduct an empirical analysis of the data using a logistic regression approach to find the most important determinants of blockholdership from a company valuation perspective. While this analysis is still limited in scope, the results already offer important insights and contribute to the literature by pointing out different characteristics of companies with or without certain blockholders. Blockholdership is extensively discussed in the academic literature (see e.g. Shleifer and Vishny, 1986; Holderness, 2003; Cronqvist and Fahlenbrach, 2009; Clifford and Lindsey, 2016; Edmans and Holderness, 2017; Backus et al., 2019, and others) while only a recent paper by Schwartz-Ziv and Volkova (2020) utilizes data sourced directly from Form 13D and 13G filings (more details and a brief review on the literature can be found in chapter 5.2).

My results, presented in chapter 5.5, show that medium-sized companies with low price-to-revenue ratios and comparatively higher equity-ratios that pay out below-average dividends are more likely to be the subject of a blockholder filing than others. These findings support the results of Hadlock and Schwartz-Ziv (2018). Holders of larger ownership percentages prefer companies with even higher equity-ratios than other blockholders and results for common valuations metrics are generally similar for non-financial blockholders compared to financial blockholders. Overall, blockholders seem to prefer companies that are moderately valued. I find no significant relationship

[8] See e.g. the white paper of the newly-founded *Long-Term Stock Exchange*, found at `https://longtermstockexchange.com/static/principals_for_lt_success_white_paper-52e153e1e0be49bd178f74475f274ef0.pdf` for an overview of literature on this topic.

of the presence of blockholders to the amount of Goodwill on the balance sheet of a company.

A revised excerpt from "Determinants of Blockholdership" was accepted for publication and is forthcoming in the Journal of Economics as „A Dataset for Blockholders in US-Listed Firms“. This article is accessible at `https://doi.org/10.1515/jbnst-2021-0033` *and the accompanying dataset can be found at* `https://doi.org/10.7910/DVN/61Z64Q`.

Value by Design? *

Stephan Kessler [†] Bernd Scherer [‡] [§] Jan Philipp Harries [¶]

September 2019

Abstract

Although academics and practitioners frequently refer in their work to equity value investing, no consensus exists as to what this style exactly encompasses. For a wide range of 3,168 alternative implementations (design choices) of what could all constitute value portfolios, we document the impact of parameter pertubations on risk-adjusted returns. The observed dispersion in Sharpe ratios allows us to identify the hierarchy of design choices and to better assess the degrees of freedom consumed within the strategy development process. We can therefore derive critical t-values that adjust for overfitting. This will prove to be useful in research governance and strategy selection.

Keywords: Risk Premia, Portfolio Construction, Value Factor, Data Snooping

JEL Codes: C63, G11, G12

This article has been published in the Journal of Portfoliomanagement Quantitative Special Issue 2020, 46 (2) 25-43, ISSN 0095-4918. It is not contained in the published version of this dissertation due to copyright reasons but can be accessed at `https://doi.org/10.3905/jpm.2019.1.122`.

*This paper has been awarded the 2020 EQDerivatives Award for the Systematic Investing Research Paper of The Year. We are grateful for all comments received at the 2019 London Quant Group Seminar in Cambridge. In particular, we want to thank Lester Ashly, Ed Fishwick, and Robert McGrabbe.

[†] University of St. Gallen - Swiss Institute of Banking and Finance, Dufourstrasse 50, 9000 St.Gallen, Switzerland

[‡] EDHEC, 24 Avenue Gustave Delory, 59100 Roubaix, France

[§] Corresponding Author

[¶] University of Wuppertal, Gaußstraße 20, 42119 Wuppertal, Germany

The Choice
Reviewing the Impact of Modeling Choices on the Fama and French Three-Factor-Model

Jan Philipp Harries [*] Stephan Kessler [†] Bernd Scherer [‡] [§]

This version: May 2021

Abstract

We investigate the importance of implementation choices for asset pricing factors using Fama and French's Three-Factor model. Our analysis concentrates on the evaluation of seemingly innocent choices around universe selection, breakpoint selection, rebalancing frequency, weighting decisions and others. These choices create significant dispersion in factor model asset pricing performance that cannot be ignored as second-order implementation noise. Our results indicate that factors should not be evaluated independently from design choices. The variation of design choices also reveals a substantial trade-off between the performance of different factors within a model which cannot be disregarded when comparing factors.

Keywords: Replication Study, Factor Fishing, Persistence

JEL Codes: B26, G11, G12

[*] University of Wuppertal, Gaußstraße 20, 42119 Wuppertal, Germany

[†] University of St. Gallen - Swiss Institute of Banking and Finance, Dufourstrasse 50, 9000 St.Gallen, Switzerland

[‡] EDHEC, 24 Avenue Gustave Delory, 59100 Roubaix, France

[§] Corresponding Author

3.1 Introduction

The Three-Factor model (hereafter referred to as FF3F) by Fama and French (1993) provides the foundation of modern asset pricing. This model was among the first that empirically attributed cross-sectional stock returns to multiple, economically motivated tradeable and priced risk factors. Since its original publication in 1993, much attention in the asset pricing literature centers around the question how good the model really works for asset pricing. Particular focus is on the development of alternative models or additional factors that claim to improve upon the FF3F and its explanatory power for the cross-section of stock returns. Recent examples in the literature include, for example, Hou et al. (2019), Kan et al. (2019) or Hanauer (2020). Fama and French (2015) themselves contributed to this growing literature by publishing an expanded five-factor model. Given the enormous popularity even 30 years after publication, we take a second look at the inner workings of the FF3F focusing on neglected research questions around factor design.

While the FF3F and other factor choices have been thoroughly analyzed in the literature[1], there is little attention paid to the more subtle design choices made during the process of calculating factor returns from raw fundamental and stock price data. In fact, while the economic reasoning and data behind the factors is heavily debated since the original publication, the less prominent design choices of FF3F are under surprisingly little scrutiny and often applied unquestioned and inconsistently to new factors. From the choice of the applicable universe at the start of the factor creation process to weighting and winsorizing choices at the end, there are many calculation steps that require decisions to be made, even if these decisions were made unconsciously at first. In their paper from 1993, Fama and French did not motivate most of their design choices. Instead they state: "The hope is that the tests here and in Fama and French (1995)[2] are not sensitive to these choices. We see no

[1] Compare e.g. Harvey et al. (2016) with a focus on datasnooping tests, Fama and French (2018) which highlights multiple comparisons issues or, more recently, Hou et al. (2020) who document the impact of data mining in factor research with a detailed performance analysis of an extensive anomaly database.

[2] First published in 1992 as working paper with the title "The economic fundamentals of size and book-to-market equity" which was later changed to "Size and book-to-market factors in Earnings and Returns" for the revised Journal of Finance version in 1995.

reason to argue that they are."[3]

However since then, evidence for the opposite case emerged. At least since Gelman and Loken (2013) introduced their "garden of forking paths", researchers increasingly became aware of multiple comparison problems due to concious or inconcious choices in research design and that results are most often indeed sensitive to these choices. Huntington-Klein et al. (2021) for example instructed seven scientists to replicate two published causal empirical results in applied microeconomics and found that the standard deviation of estimates across replications was 3–4 times the mean reported standard error due to not-documented researcher choices. In empirical finance, Asness and Frazzini (2013) already showed that seemingly innocuous choices like the use of a more recent price in the calculation of the book-to-price ratio can cause a huge difference in results. Harvey (2017) used an highly-significant but nonsensical factor[4] as an empirical example, demonstrating that many empirical design choices may be crucial for the results and argued for more replication efforts.

Besides the question of how sensitive the results are to choices, the other interesting research question is how good the original FF3F specification's (hereafter called *the choice*) ability to price assets is compared to other reasonable specifications. Are design choices of second-order importance or do they induce considerable variation in single factor and factor model performance? Is there a hierarchy of choices?

Our contribution to the literature is threefold. First, we analyze in detail, how these subtle and seemingly unimportant design choices influence the performance and explanatory power of the FF3F. We study whether minor differences in factor design would significantly change the results despite an unchanged economic foundation. If design choices create additional and sizable return dispersion, they need to become part of the the return stream permutations used for multiple hypothesis tests as they can be equally tempting to use for data mining. As we will show in section 3.4, dispersion caused by design choices does not affect all factors equally. Second, we look at the many design choices Fama and French could have taken using

[3] Their predating paper ("The Cross-Section of Expected Stock Returns" - 1992), which introduces the size (SMB) and book-to-market (HML) factors after providing evidence inconsistent with the capital asset pricing model (CAPM) by Sharpe (1964), Lintner (1965) and Black (1972) for describing the cross-section of expected returns, explains some, but not all important design choices.

[4] Based on letters present in the ticker symbol.

only information they had at the time of *the choice* and evaluate them using out-of-sample information. Here, we are less interested in post publication decay in factor returns (compare, e.g., Mclean and Pontiff, 2016, who found that performance of systematic factors deteriorates after publication) but in the persistence of in-sample design fishing. Would Fama and French arrive at the same design choices using only out-of-sample (post publication research) data? With 2,592 permutations of design choices[5] at our hand, we can also evaluate whether Fama and French (1993) lost important degrees of freedom through their design choices which decreased the pricing power of their model. Third, we are also interested in the investor's point of view. How does *the choice* affect the ability of investors to achieve higher risk-adjusted returns (from multi-factor investing) versus returns from a single market factor (CAPM) model?

Our analysis finds large dispersion in the risk-return characteristics of the individual factors for alternative implementation choices. Interestingly, we find a significant negative relationship between single-factor performance across design choices. Consequently, maximum Sharpe ratios for factor models are less sensitive to design choices than maximum Sharpe ratios for single factors. In the case of the FF3F, choices that lead to an improved SMB performance mostly result in an inferior HML performance and vice versa. This indicates that not only the relative performance of factors in a factor model is affected by design choices but also hints at a comparison issue when testing new factors or anomalies with existing factor models where choices are fixed by convention. The original FF3F choices do not lead to superior risk-adjusted returns when compared to other design options, neither in- nor out-of-sample. This is true for a single factor or for the combined factor model. Our results indicate that performance,as measured by Sharpe ratio, was not a main driver of the design choices made by Fama and French. We find that universe selection, breakpoint selection as well as asset weighting are important choices. When added jointly to the market portfolio, Size and Value factors turn out to be economically meaningful yielding an efficient frontier with higher risk-adjusted returns independent from design choices but with less than

[5] The 2,592 permutations of design choices lead to 2,592 Three-Factor models whose return timeseries are analyzed throughout this article. These factor models and/or factor return timeseries are called implementations or variations hereafter.

the desired statistical significance (results do not exceed the 95% confidence level). From an asset pricing perspective, we find that in-sample *the choice* does well with a lower pricing error than 84% of alternative implementations. Out-of-sample, the pricing capability of all variations deteriorates considerably. While the FF3F is no longer superior, neither are its peers. We do not find variations that reliably enable significantly better, economically relevant pricing capability than the original FF3F.

A few related topics are explicitly kept out of scope in this article: We assume that the signals used to construct the model are given and do not stray into alternative factor definitions (compare Kessler et al., 2020, for a comparison of different value factor definitions). Furthermore, we do not evaluate alternative factor models due the fact that no alternative model enjoys the same influence on empirical asset pricing as the FF3F. Additionally, the relatively shorter out-of-sample periods for newer models (such as the five-factor model introduced in 2015 by Fama and French) pose a challenge in the evaluation of their robustness. We want to obtain general results for the "gold standard", leaving further work on specific models to others overtime.

The remainder of this article is structured as follows: In Section 3.2, we outline the literature summarizing the development of the FF3F, the history of choices for the model and related approaches at dissecting its performance. Afterwards in Section 3.3, we give a brief overview on the data we use and a detailed description of the actual permutations of design choices applied in our study. Section 3.4 focuses on the characteristics of the different factor implementations for SMB and HML, as well as the key drivers behind (pricing) performance differences. We then turn towards the evaluation of the alternative implementations of the Three-Factor model from an investor's perspective in Section 3.5 before we turn to an asset pricing context in Section 3.6. Section 3.7 concludes.

3.2 History of Choices and Literature Review

When Fama and French (1993) constructed the Three-Factor model, they faced a range of relevant design choices dealing with questions about the scope of the universe, timing for data updates and merging as well as portfolio construction. In this section, we review some of the most relevant choices made by them and the literature related to these choices to establish a common ground before introducing our variations in the next section. In doing that, we don't imply that all discussed choices are of equal importance or even that all choices were made consciously at the time of publication.

As mentioned in Section 3.1, Fama and French introduced some of their design choices in the preceding article from 1992, "The Cross-Section of Expected Stock Returns". There, the 6-18 month data availability delay (related to our *merge_date*[6] variation) before newly published book value data is used in the formation of BE/ME-based portfolios is called "conservative." It is motivated with the results of Alford et al. (1994), who find that 20 percent of 10-Ks are filed after the 90-day statutory due date and that late-reporting firms are not a random sample, but experiencing more unfavorable economic events on average. However, the overwhelming majority of late reports is delayed less than 30 days and with SEC rule 33-8644 which came into effect in 2005, the 10-K filing period was reduced from 90 to 60 (or 75) days for more recent data. Earlier work, e.g. Basu (1983), already assumed accounting data to be available within three months of fiscal year-ends. Additionally, Easton and Zmijewski (1993) find that on average, the Earnings Announcement Date after which up-to-date book-value data should be assumed as available is 44 days and 38 days for NYSE/AMEX and NASDAQ, respectively, earlier than the 10-k filing. Thus, an assumed publication delay of three months after fiscal year-end could arguably have been a more reasonable choice than the availability of new data in June of the following year, which results in the aforementioned variable data availability delay of 6-18 months. Fama and French write that tests using a smaller sample of firms with only December fiscal year-ends, omitting the variable delay length (which is fixed by design in our "three months" variation), yielded similar results compared to their primary methodology. Retrospectively, it seems reasonable to choose the more conservative publication lag to avoid any potential look-ahead bias. Nevertheless a shortened lag would result in a factor calculation that is closer to its economic foundations and could prevent misinterpretations.

[6] See Section 3.3 for a detailed description.

For the denominator of BE/ME, Fama and French (1992) mention that their “use of December market equity (...) is objectionable for firms that do not have December fiscal year-ends”, as numerator and denominator of BE/ME are not aligned in this case. However, using ME at fiscal year-ends could introduce cross-sectional variations in the measurements if there are significant market-wide changes in stock prices during the year. They report that the use of fiscal-year-end ME’s has little impact on their return tests which we test with our *bookvalue_date* variation.

Fama and French’s (1993) paper “Common risk factors in the returns on stocks and bonds”, which is generally viewed as the source of the original Three-Factor Model, expanded on the preceding findings and introduced a new methodology. Instead of using cross-sectional Fama and MacBeth (1973) regressions on single characteristics (such as ME/BE), Fama and French introduced the time-series regressions using excess returns or returns on zero-investment portfolios as explanatory variables, which have become standard in asset pricing. The 2x3 sorts for the six Size and Book-to-Market portfolios and the NYSE-only portfolio breakpoints were first introduced in Fama and French (1995) without further explanation and motivation.[7] We assume that the Median split for Size and the 30/40/30 (also called “Wing”) split for BE/ME were probably related to data availability. For 3x3 sorts or a narrower split (e.g., 20/60/20), some portfolios become very small and contain only few stocks in the earlier part of the data. Fama and French (1993) themselves state, that they use three groups on BE/ME due to its stronger role compared to Size and note that “the splits are arbitrary, however, and we have not searched over alternatives”. In their later Five-Factor Model (Fama and French, 2015), they also used 2x2 splits. In more recent literature, more pronounced splits have become popular. Stambaugh and Yuan (2017), e.g., use the 20th and 80th percentiles of all NYSE/AMEX/NASDAQ stocks in the calculations for their *q-model*, noting that “relative mispricing in the cross-section is likely to be more a property of the extremes than of the middle.” Goyal (2019) notes that there is “nothing magical about a 30-70 or a 20-80 split” and that many choices could be driven by data considerations and calls for future work to address these issues. In his words, “all methods of factor construction are

[7] Please note that we don’t have access to the first working paper version of that paper which was first published in 1992 and originally quoted in Fama and French (1993).

ad-hoc." We vary these choices in our *mcap_percentiles* and *beme_percentiles* parameters.

The use of NYSE breakpoints increased continuity as otherwise breakpoints would have jumped considerably after the inclusion of AMEX and NASDAQ stocks into the sample. Additionally, the addition of thousands of tiny stock would have significantly influenced the distribution of book-to-market ratios diminishing the influence of the stocks with the biggest share in total market capitalization. However this restriction is not uncontroversial. Bali et al. (2011) e.g. write that the calculation of portfolio breakpoints using only stocks listed on the NYSE substantially diminishes the detected magnitude of the returns associated with size investing. This is also supported by Fama and French (2018) themselves, finding that all return spreads are higher for small-cap-only factors. To understand the magnitude of this effect, we also include breakpoints from sorts of NYSE-Amex-NASDAQ stocks instead of only NYSE stocks (*only_nyse_bps*) and, for the opposite case, results from a universe restricted to only NYSE stocks (*only_nyse_stocks*) in our variations.

While Fama and French's choice of one yearly portfolio rebalancing in July was reasonable at the time and Fama and MacBeth (1973) earlier noted a necessary "balance of computation costs against the desire to reform portfolios frequently" for their 4-year testing periods, Rosenberg et al. (1985) already successfully employed monthly portfolio rebalancing, additionally stating that "trading costs (...) would almost certainly have had a negligible effect upon performance". Newer research is ambiguous about the optimal frequency of rebalancing. He and Modest (1995) and Luttmer (1996) found that transaction costs could play a significant role explaining equity premiums depending on portfolio rebalancing frequency. Lynch and Balduzzi (2000) expanded on their results examining the effect of return predictability on portfolio rebalancing choices. Smith and Desormeau (2006) find that longer time intervals for rebalancing outperform shorter periods while Almadi et al. (2014) results indicate that monthly rebalancing is superior in the absence of very high unit transaction costs and newer factor models (e.g. the q-factor model by Hou et al., 2015) also frequently use monthly rebalancing for some of their portfolios. The question of an optimal rebalancing schedule remains open, motivating us to use varying intervals described by our *rebalancing_frequency* variation.

The use of value-weighted returns on the portfolios (whose returns are then added equal-weighted) can also be questioned. Firstly, while bigger companies have indeed a bigger impact on market returns, the splits can distort the overall picture. For example, the biggest company in a low-size portfolio will have a much bigger influence on factor returns than the smallest company of a high-size portfolio, even when post companies are almost identically sized. Secondly, some studies show "that funds tend to be equally weighted to a greater degree than they are value weighted" (Block and French, 2002). Plyakha et al. (2014) show “that the inferences drawn from tests of asset-pricing models are substantially different depending on whether one uses equal- or value-weighted test portfolios”. Hou et al. (2020) argue against equal-weighting when replicating anomalies due to higher transaction costs and other problems with microcaps like microstructure frictions, bid-ask spreads and non-synchronous trading which can bias cross-sectional returns upwards. However, one could argue that trading costs are steadily diminishing while liquidity has risen since inception of the FF3F, making systematic small-cap investments more feasible. Additionally (as already discussed above), empirical results have repeatedly shown that small-cap factors have shown superior performance even when explaining large-cap portfolios. Carhart’s (1997) momentum factor, for example, is equal-weighted (and rebalanced monthly) and Fama and French themselves included equal-weighted portfolios in their earlier papers which we also do with our *weighting* variable.

Finally, the correct application of winsorization is also a controversial, albeit not much-discussed, topic in asset pricing research. Bali et al. (2016) concede that it is difficult to decide when to use winsorization as some outliers are legitimate while others are only data errors. From an investment perspective, winsorization is not a sensible choice as winsorized returns are not investible per se (albeit it could be possible in theory to achieve some degree of winsorization with options). In earlier research, Bali et al. (2011) apply winsorization themselves while at the same time finding that extreme positive returns have a significant effect in the cross-sectional pricing of stocks. Winsorization is also commonly applied to check raw results for robustness (see, e.g., Brennan and Wang, 2007). While it is one of the great advantages of the asset pricing approach to economical questions that winsorization is not necessarily needed and most outliers in the CRSP data are believed to be real, we try to get a better understanding of the influence of extreme returns on factors with our *winsorization* variation.

It is instructive to see how the rich literature related to the Fama and French (1993) model raises a large range of questions around the design choices. However, as far as we know, there doesn't yet exist a coherent analysis which evaluates those choices and finds conclusive answers as to which design choices are the most important ones in an asset pricing context. Previous work focused either on the specific choices that led to mispricing (like e.g. Cremers, 2013, who analyses the reason for nonzero alphas when pricing benchmark indices and proposes modified boundaries between size and value groups and using value-weighting for SMB and HML calculation with the 2x3 size/value portfolios) or on the degree and effects of data mining and dredging for asset pricing research (like e.g. Hou et al., 2020, with their replication and analysis of more than 400 published anomalies with comparable design choices or Yan and Zheng, 2017, who built and analyzed an universe of 18.000 permutations of fundamental signals). In the following section, we outline our framework and describe the required data and calculation of variations to address those questions.

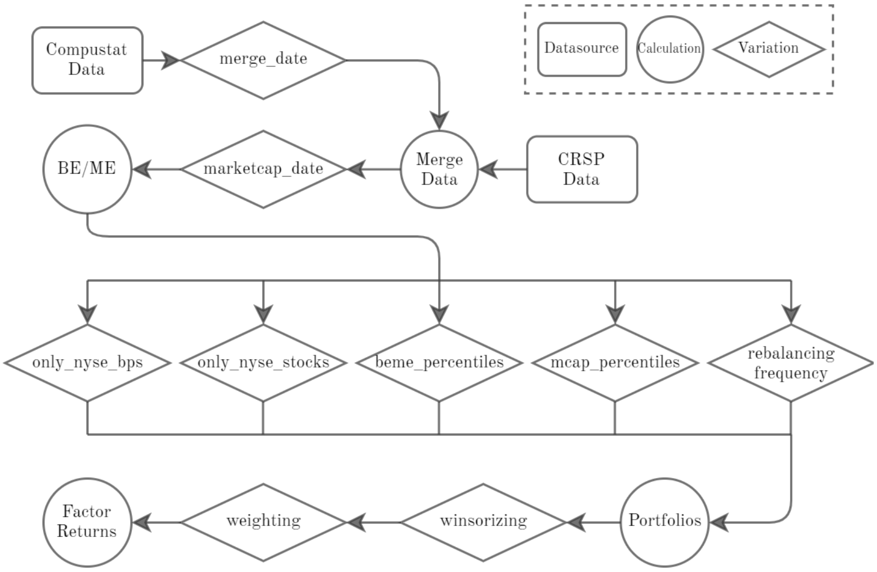
3.3 Data and Variations

Following Fama and French (1993), we use data on all NYSE, AMEX and NASDAQ stocks with share codes 10 or 11 from both CRSP and Compustat for our calculation of *Size* and *BE/ME*. Consistent with Fama and French (1993) our first sample period ("in-sample") ranges from July 1963 to December 1991, while our second sample period ("out-of-sample") starts in January 1992 and ends in December 2020. The number of stocks in our sample is the same as in the original paper (Fama and French, 1993) as we use the same drop-out criteria. Benchmark factors and the one-month T-bill rate were obtained from Ken French's website.

Our portfolio-building and factor-creation process follows Fama and French's but includes variations of nine important design parameters at different stages and is illustrated in Figure

Figure 3.1: **Factor-creation Process**

This figure illustrates our process of calculating the portfolios and factor return time-series for the design choice-induced variations of the Three-Factor Model. Rectangles denote data sources, circles denote calculation steps and rhombuses denote our variations.



3.1.

Yearly book common equity, which is defined like in previous literature as the book value of stockholder's equity plus balance-sheet deferred taxes and investment tax credit minus the book value of preferred stock[8] is taken from the Compustat database for all companies, which have appeared on Compustat for at least two years. [9].

Our first variation, *merge_date*, changes the date, at which stock price data and book equity data from CRSP and Compustat are merged together. In other words, *merge_date* describes the date, at which information about a company's book value can be used to sort it into a portfolio. The original *merge_date* is *next_june*, which means book value data reported in the annual report of year t (which mostly refers to the business year ending

[8] As in Fama and French (1995), redemption, liquidation or par value (in that order) is used to estimate the value of preferred stock.

[9] Some newer factor models like e.g., the "q"-factor model (Hou et al., 2015) use quarterly Compustat earnings and balance sheet data as well. In non-reported results, we also use quarterly Compustat data for BE/ME calculation, which is available starting in 1972, as an additional variation and robustness check. The general results are comparable but we exclude them from reported results due to the not-comparable availability time-frame and different data fields.

at December 31th, but also can be in January for some companies which would imply a publication delay of 17 months) is available for portfolio sorts in June of year $t+1$ and can only be used without ambiguity for yearly Compustat data. Alternatively, *merge_date* can be *3_months*, which means book value data is assumed to be available three months after the respective reporting date (e.g., on March 31th of year $t+1$ if the business year ends on December 31th of year t). This choice for the variation is founded in the observation that most reports are published 2-3 months after the reporting date (see also 3.2) and thus can be used at latest 3 months later.

Secondly, we vary the *marketcap_date* for the denominator of *B/M.* Fama and French use the market capitalization at the end of year t for *B/M* calculation in June of year $t+1$, which means there can be up to 11 months between the reporting date of book value and the date when market capitalization is measured. We add *be_time*, which synchronizes the date for both sides of *B/M.* One advantage of a unified date to measure the market capitalization like in *prev_dec* is improved cross-sectional comparability; however in this case one could also argue for the use of the latest available stock price data (e.g., in the original case not December of year t but June of year $t+1$), like Asness and Frazzini (2013) did.

After applying the first three variations, we get 6 variants of a monthly dataset with pre-calculated values for *Size* (market capitalization) and *BE/ME* (book-to-market) for sorting and portfolio building. The next five variations are applied in parallel before breakpoints are calculated and the stocks are sorted into portfolios.

The third variation, *only_nyse_bps*, controls whether only NYSE-listed companies are included in the calculation of breakpoints (as in Fama and French's original paper) or not. While this choice made sense at a time when most non-NYSE stocks were too small to be traded actively, we argue that due to total market growth and improved transparency, there is no economic reason anymore to artificially restrict the universe of stocks from which breakpoints are calculated[10]. To be consistent in design choices, we also add a fifth variation, *only_nyse_stocks*, where only NYSE-listed stocks are included in the portfolios for the default case where only NYSE-listed stocks included in the breakpoint calculation.

[10] Note, that our data only includes NYSE, NASDAQ and AMEX stocks by default.

The fifth and sixth variations, *beme_percentile* and *mcap_percentile*, contain different percentiles, at which the breakpoints for the independent sorts are set. Fama and French use one breakpoint at the Median (50th percentile) for market capitalization (size) and two breakpoints at the 30th and 70th percentile (also called Wing Portfolios) for *BE/ME* to create 6 intersection portfolios from the resulting independent 2x3 sorts. To our knowledge, there are no economic reasons given for these choices and they seem somewhat arbitrary. We vary *mcap_percentile* to the 30th and 70th percentile as well (which results in independent 3x3 sorts with 9 intersection portfolios) and vary *beme_percentile* to the 50th (2x2/3x2) and 20th and 80th percentile alternatively.

The seventh variation is the *rebalancing_frequency*[11], for which we implement different rebalancing and portfolio building regimes. In the default case, sorts are conducted and portfolios are built once every year at the end end of June and stocks are held held from July of year t until the end of June of $t + 1$. Additional choices are quarterly (end of March, June, September and December) and monthly sorts and rebalancing.

After companies are sorted in the resulting intersection portfolios, the final two variations affect how portfolio performance is calculated. Our eighth variation is cross-sectional *winsorization* of stock returns at the 1% and 99% percentiles for each months, as commonly used in economic studies. The ninth and final variation refers to the *weighting* of stocks in the portfolios. While portfolio returns are value-weighted by default, we also include equal-weighted portfolios[12] in our calculations.

Following Fama and French (1993), we get 4-9 timeseries of portfolio returns (in the default case the 6 portfolios *BH*, *BL*, *MH*, *ML*, *SH* and *SL*). The *SMB* factor is then calculated by subtracting the average return of the *B** portfolios from the average return of the *S** portfolios and the *HML* factor by subtracting the average return of **L* portfolios from the average return of the **H* portfolios. All variations combined, we arrive 2,592 triples of MKT, SMB and HML factor timeseries which we then use for our following analysis.

[11] Please note that we implement the portfolio rebalancing independently from data availability. For additional information we refer to the explanation for the *merge_date* variations.

[12] Note that the weighting is updated according to *rebalancing_frequency*, independent of the availability of new data.

3.4 Design Choices and Their Impact on Factor Characteristics

This section documents the impact of factor design choices on the risk and return characteristics of individual factors. Fama (1991) himself conjectured in “Efficient Capital Markets: II” that a large part of return predictability in academic studies due to the effect of data mining. Mclean and Pontiff (2016) find that the average predictor’s return declines by 58% post-publication which they attribute to both, statistical biases and price impact of sophisticated traders after publication.

While they included factors and characteristics with diverse economic foundations and unified some design choices,[13] our research question leads us in the opposite direction. We leave factor definitions the same and deliberately vary implementation choices (we use implementation and design interchangeably). If design choices play only a minor role, we would expect that post-publication performance metrics should be similar for all variations. To be able to compare factor performance from before and after publication of Fama and French’s original model, we define the period from July 1963 to December 1991 as in-sample period and the period from January 1992 to December 2020 as out-of-sample period. Summary statistics (based on monthly return series) for all design choices - described in Section 3.3 - are given in Table 3.1. We find that risk-return characteristics derived from the return series for our design choices display substantial variation. Average in-sample Sharpe ratios for the value factor (HML) are 0.56, ranging between 0.22 and 1.01. For the out-of-sample period, we find slightly smaller dispersion with Sharpe ratios between -0.11 and 0.65, averaging 0.24. We also observe a similar dispersion in Sharpe ratios for the size (SMB) factor, albeit at the expected lower absolute level, ranging from -0.13 to 0.46 (with an average of 0.22) in-sample and -0.51 to 0.37 (average 0.08) out-of-sample. The HML returns of Fama and French’s original model are close to the mean returns of all our variations in-sample but perform worse out-of-sample (4.59% vs 4.74% in-sample and 1.46 vs 2.82% out-of-sample). For SMB, the Fama and French version exhibits similar in-sample returns (3.16% vs 2.54%)

[13] By creating long-short return timeseries with Median-split portfolios for each factor.

and lower absolute but higher relative (to the mean of all variations) out-of-sample returns (1.70% vs 1.01%). Notably, HML and SMB perform worse on average in the out-of-sample period compared to the in-sample period in almost all metrics. This is particularly true for SMB which experiences a significant drop in performance as shown by a drop of almost two thirds in average Sharpe ratio.

From an asset pricing perspective, it is also noteworthy to highlight that the risk characteristics are rather sensitive to modeling choices. For example, the in-sample volatility of the HML factors ranges from 5.55% to 11.27% with similarly broad dispersion found for the SMB factors (8.09% to 16.29%). Risk characteristics measured through skewness, kurtosis and maximum drawdowns vary substantially across the variations, indicating that the actual nature of the risk premium captured by the factors is sensitive to implementation choices. Out-of-sample factor returns also show higher average out-of-sample volatility with maximum HML volatility rising from 11.27% to 16.35%. This will make it more difficult to find statistical significance in in-sample and out-of-sample comparisons.

Table 3.1: **Summary Statistics for Factor Performance**

Descriptive summary statistics (MDD stands for maximum drawdown) for all 2,592 design variations for in-sample (1963:07 to 1991:12) and out-of-sample (1992:01 to 2020:12) for HML and SMB factor. FF3F denotes values for the original FF design choice.

		In-sample						Out-of-sample					
		FF3F	mean	median	stdev	min	max	FF3F	mean	median	stdev	min	max
HML	Average Return	4.59%	4.74%	4.27%	1.90%	1.65%	10.15%	1.46%	2.82%	1.95%	2.49%	-0.64%	9.91%
	Volatility	8.89%	8.36%	8.56%	1.64%	5.55%	11.27%	10.85%	10.39%	9.96%	2.66%	5.29%	16.35%
	Sharpe ratio	0.52	0.56	0.53	0.18	0.22	1.01	0.13	0.24	0.20	0.19	-0.11	0.65
	Skewness	-0.02	-0.13	-0.13	0.13	-0.53	0.35	0.07	-0.12	-0.11	0.27	-1.09	0.65
	Kurtosis	1.26	1.37	1.30	0.62	0.35	7.74	2.60	3.96	3.64	2.05	1.03	12.92
	Min	-9.99%	-9.32%	-9.25%	2.24%	-20.63%	-5.29%	-13.96%	-14.85%	-14.18%	5.06%	-32.50%	-6.48%
	Max	8.65%	8.27%	8.00%	2.59%	4.28%	18.40%	12.58%	12.51%	12.62%	3.94%	4.73%	22.27%
	MDD	-28.44%	-28.91%	-29.33%	4.62%	-41.19%	-16.80%	-58.68%	-47.86%	-49.19%	7.48%	-63.72%	-26.90%
	MDD / Volatility	-3.20	-3.52	-3.49	0.49	-4.88	-2.40	-5.41	-4.85	-4.60	1.19	-8.49	-2.99
SMB	Average Return	3.16%	2.54%	2.82%	1.39%	-1.37%	6.33%	1.70%	1.01%	1.57%	2.33%	-8.02%	6.16%
	Volatility	9.97%	12.07%	12.48%	2.15%	8.09%	16.29%	11.11%	13.13%	13.09%	2.28%	8.42%	17.72%
	Sharpe ratio	0.32	0.22	0.25	0.12	-0.13	0.46	0.15	0.08	0.13	0.17	-0.51	0.37
	Skewness	0.15	0.40	0.42	0.17	-0.03	0.88	0.64	0.98	1.08	0.52	-0.09	1.83
	Kurtosis	1.13	1.69	1.41	0.73	0.53	4.06	7.03	6.55	7.32	3.43	0.50	13.98
	Min	-9.88%	-11.25%	-11.38%	1.82%	-14.97%	-7.59%	-16.82%	-15.38%	-15.71%	3.98%	-23.75%	-7.80%
	Max	11.01%	14.03%	13.78%	2.51%	9.20%	21.64%	21.19%	22.72%	23.11%	6.96%	8.39%	35.62%
	MDD	-46.51%	-60.43%	-59.60%	11.28%	-85.29%	-38.38%	-41.80%	-51.73%	-47.17%	14.55%	-95.61%	-28.04%
	MDD / Volatility	-4.67	-5.03	-4.90	0.58	-7.18	-3.62	-3.76	-3.96	-3.65	0.92	-7.49	-2.69

Table 3.2: **Correlation across Factors and Sample Periods**

The table shows the distribution of pairwise correlation between design choices within SMB and HML denoted by $\rho\left(SMB_i, SMB_j\right)$and $\rho\left(HML_i, HML_j\right)$as well as between factors, but for the same design choice.

	$\rho\left(SMB_i, SMB_j\right)$		$\rho\left(HML_i, HML_j\right)$		$\rho\left(SMB_i, HML_i\right)$	
Percentile	1963:7 to 1991:12	1992:1 to 2020:12	1963:7 to 1991:12	1992:1 to 2020:12	1963:7 to 1991:12	1992:1 to 2020:12
1%	0.624	0.194	0.723	0.531	-0.201	-0.507
5%	0.693	0.325	0.767	0.587	-0.182	-0.482
25%	0.854	0.634	0.841	0.695	-0.125	-0.426
50%	0.912	0.805	0.889	0.833	-0.085	-0.360
75%	0.951	0.893	0.920	0.901	-0.051	0.093
95%	0.987	0.978	0.960	0.959	0.020	0.236
99%	0.996	0.995	0.986	0.987	0.045	0.286

Table 3.2 shows summary statistics for pairwise return correlations of the value factor and size factor implementations, yielding additional evidence that the design choices can be highly impactful for the return characteristics of a factor. For SMB factors in the in-sample period, we find a median pairwise correlation of 91%. The 5th and 95th percentile of realized pairwise correlations are 69% and 99%, respectively. The median pairwise correlation among the HML implementations is with 89% marginally lower than for SMB, while 5th and 95th percentiles for HML are slightly narrower with 77% and 96% respectively. In the out-of-sample period, pairwise correlations are substantially lower. Correlations at the bottom percentile go down from 62% to only 19% for SMB and from 72% to 53% for HML. The correlation between SMB and HML for the same design choice is mostly negative and widens out-of-sample. This indicates again that the large impact of some - seemingly minor - implementation details was not obvious at time of publication and that it is almost impossible to account for these deviations ex-ante, which poses the challenging question whether any single factor model can deliver on its promises or we would actually need a portfolio of related models based on the same economic premise but spanning the space of reasonable implementation choices.

In Table 3.3 we sort all 2,592 SMB design variations into decile buckets by their in-sample performance (measured as the Sharpe ratio) and display the corresponding returns at the 1st, 50th and 99th percentile for each decile portfolio for in-sample and out-of-sample periods. As can be seen, the Sharpe ratios are substantially smaller in the out-of-sample period for all decile buckets, both in the median and in the extremes. Noteworthy, the difference between in-sample and out-of-sample Sharpe ratio is even bigger for the in-sample buckets in the tails of the distribution.

Table 3.3: **Performance and Decay for SMB**

Sharpe ratios for all 2,592 design variations are sorted into 10 buckets (conditional on their in-sample performance from 1963:07 to 1991:12) from worst (0th to 10th percentile) to best (90% to 100th percentile). For each bucket we also document the variations across its constituents, i.e., we report 1st, 50th (median) and 95th percentile. Out-of-sample Sharpe ratios for realized performance from 1992:1 to 2020:12 are conditional on the in-sample placement of a given design choice into performance buckets. Finally, we calculate the percentage of factors within each decile bucket that display a statistically significant deterioration of out-of-sample performance for 95%, 97.5% and 99% confidence intervals, employing a bootstrapping procedure.

	1963:7 to 1991:12			1992:01 to 2020:12			% of design choices with statistically significant performance decay		
Percentile	1st perc	50th perc	99th perc	1st perc	50th perc	99th perc	95% conf	97.5% conf	99% conf
(worst) 0-10	-0.114	-0.011	0.045	-0.494	-0.224	-0.025	26.923	16.154	8.846
10-20	0.047	0.081	0.111	-0.255	-0.111	0.086	8.880	0.386	0.000
20-30	0.112	0.137	0.155	-0.218	0.016	0.152	11.969	1.544	0.000
30-40	0.156	0.178	0.200	-0.169	0.060	0.218	3.462	3.077	0.000
40-50	0.201	0.222	0.246	-0.043	0.085	0.254	0.000	0.000	0.000
50-60	0.247	0.262	0.280	0.015	0.156	0.284	0.000	0.000	0.000
60-70	0.280	0.293	0.303	0.071	0.178	0.299	0.000	0.000	0.000
70-80	0.304	0.314	0.323	0.098	0.200	0.306	0.000	0.000	0.000
80-90	0.323	0.333	0.352	0.118	0.217	0.335	0.000	0.000	0.000
(best) 90-100	0.353	0.374	0.447	0.131	0.272	0.365	0.000	0.000	0.000
Fama/French		0.317			0.153		no	no	no

Table 3.4: **Performance and Decay for HML**

Sharpe ratios for all 2,592 design variations are sorted into 10 buckets (conditional on their in-sample performance from 1963:7 to 1991:12) from worst (0th to 10th percentile) to best (90% to 100th percentile). For each bucket we also document the variations across its constituents, i.e., we report 1st, 50th (median) and 95th percentile. Out-of-sample Sharpe ratios for realized performance from 1992:1 to 2020:12 are conditional on the in-sample placement of a given design choice into performance buckets. Finally, we calculate the percentage of factors within each decile bucket, that display a statistically significant deterioration of out-of-sample performance for 95%, 97.5% and 99% confidence intervals, employing a bootstrapping procedure.

	1963:7 to 1991:12			1992:1 to 2020:12			% of design choices with statistically significant performance decay		
Percentile	1st perc	50th perc	99th perc	1st perc	50th perc	99th perc	95% conf.	97.5% conf.	99% conf.
(worst) 0-10	0.229	0.308	0.342	-0.098	0.068	0.164	23.846	6.538	1.154
10-20	0.343	0.367	0.386	-0.055	0.046	0.171	65.251	28.571	7.336
20-30	0.387	0.405	0.423	-0.055	0.078	0.156	78.378	36.680	15.830
30-40	0.424	0.443	0.461	-0.054	0.106	0.213	90.769	60.769	17.308
40-50	0.462	0.483	0.529	-0.009	0.104	0.207	95.367	84.170	54.440
50-60	0.543	0.610	0.637	0.184	0.267	0.474	72.973	45.174	10.425
60-70	0.639	0.664	0.687	0.210	0.324	0.504	69.615	52.692	26.154
70-80	0.688	0.709	0.736	0.252	0.376	0.590	63.707	52.896	23.938
80-90	0.737	0.780	0.827	0.309	0.474	0.628	51.351	26.641	10.039
(best) 90-100	0.831	0.872	0.990	0.407	0.539	0.636	65.769	39.615	10.000
Fama/French		0.516			0.135		yes	yes	yes

As our approach could be described by "data mining on purpose," naive measures of strategy decay in the out-of-sample period are difficult to apply and interpret. To circumvent this problem we apply a bootstrapping technique which identifies statistically significant performance decay with greater certainty and corrects for data mining. First, we bootstrap 5,000 new return series for each factor (with replacement) from the out-of-sample data. Then, we calculate the respective Sharpe ratio for each of these return series to get 5,000 Sharpe ratios for each model specification.[14] As evident in the following results, for most design choices we cannot establish a statistically significant decay after this bootstrapping procedure. While Sharpe ratios are lower in-sample, in most cases this is still consistent with pure chance. The bottom four buckets also contain variants that exhibit a significant performance decay after applying our bootstrapping procedure (at the 95% and 97.5% confidence levels), with the highest rate of decaying variants (26.9%) in the lowest bucket. This bucket is also the only one containing variants with decaying performance at the 99% confidence level. The FF3F SMB factor does not show a statistically significant decay at the 95% confidence level after our bootstrapping procedure.

For the HML factor variations, table 3.4 paints a somewhat different picture. The median factor Sharpe ratios are lower out-of-sample for all buckets. Our bootstrap analysis results in a significant share of variants with statistically significant performance decay for all tested significance levels and buckets, most pronounced in the buckets near the median. However, despite the significant decay, HML design choices with superior in-sample performance tend on average to display relatively better out-of-sample performance versus their peer factors. The original Fama French HML factor belongs to an average in-sample performance bucket with an in-sample Sharpe ratio of 0.516 but performs substantially worse out-of-sample with a Sharpe ratio of only 0.135. This under-performance is also significant after our bootstrapping procedure and could, thus, be interpreted as the result of statistical bias or over-fitting during parametrization in the design process. Most interestingly, both SMB and HML design choices show some form of persistence. Design choices that perform best (worst)

[14] Given that we work with monthly returns, no attempt (e.g., non-parametric sampling with random blocks or parametric sampling from a fitted data generating process) has been made to account for dependency as auto-regressive effects in returns, as volatility tends to be low in monthly data.

in-sample also display on average better (worse) out-of-sample performance. This provides limited evidence that design fishing[15] offers some value.

Finally, we look at the relation between Sharpe ratios for HML and SMB within the same design. Is there a trade-off among design choices, or do design choices that result in high Sharpe ratios for HML also lead to high SMB Sharpe ratios? The results are shown in Figure 3.2. The relation between HML and SMB Sharpe ratios slopes down with an R-squared of 0.42. Whatever design choice makes a good value factor, makes a poor size factor in return. This is of large practical importance as this negative correlation must have complicated the search for the best three-factor model. The FF3F *choice* sits comfortably in the middle of most choices, while we find also many variations which trade a lower SMB Sharpe ratio for a higher HML Sharpe ratio. This trade-off could also affect other, new factors and anomalies when measuring them against benchmark models like the FF3F. It is plausible that factor performance (and the significance) depends on the underlying design choices as it does for SMB and HML. As far as we know, this comparison problem hasn't been addressed in the literature until now and thus demands further research.

As expected, some choices appear to be more important than others as design choices seem to cluster in risk and return space. What drives the dispersion in Sharpe ratios across our design choices? What are the most important design choices? How do design choices interact? To answer these questions, we build regression trees with the in-sample Sharpe ratios of all 2,592 individual factor variations in the cross section as dependent variable and a 2592-by-12 matrix of categorical data as explanatory variables. In this context, regression trees offer many advantages over linear regressions. The first split selects the most important variable while the sequence of splits is able to model nonlinearities which allows us to find otherwise hidden nonlinear interactions. While linear regressions can also uncover nonlinear interactions by including all possible cross terms, this requires as many right-hand-side variables as data points and thus results in a loss of all degrees of freedom. Most importantly the sequential and interactive nature of regression trees directly mimics the decision taking

[15] Cochrane (1996) coined the term factor fishing, i.e., the selection of factors on the grounds of purely statistical information. We build on this terminology to coin the term design fishing as the selection of factor designs based purely on statistical information.

Figure 3.2: **One Size fits All?**

Plot of full-sample HML Sharpe ratios versus full-sample SMB Sharpe ratios with fitted (OLS) regression line $\widehat{SR}_{HML} = 0.50 - 0.82\widehat{SR}_{SMB}$. Each data point reflects one design choice. The FF3F *choice* is marked red.

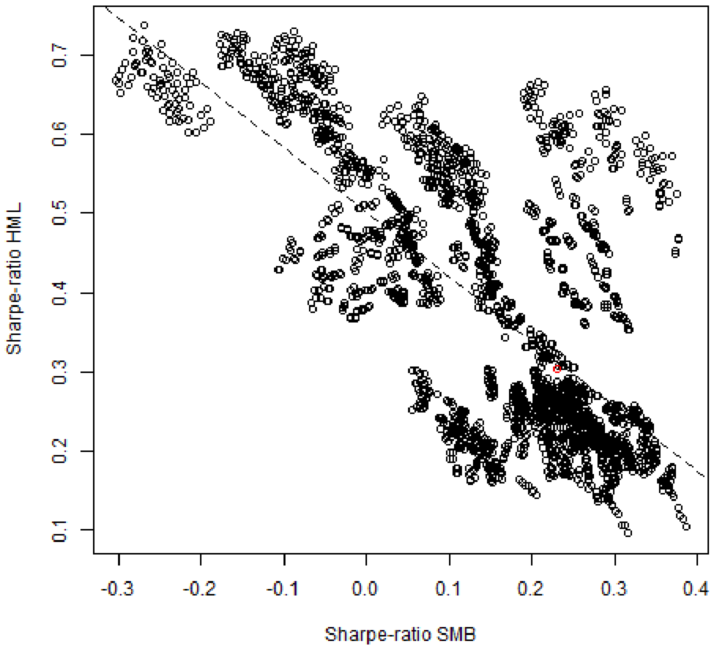
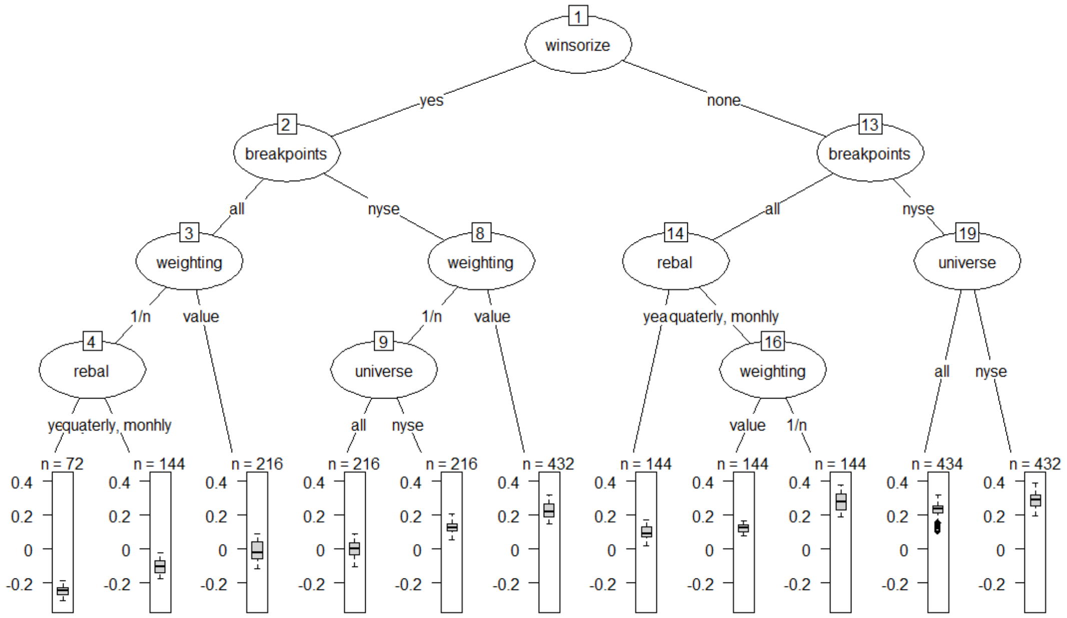


Figure 3.3: **SMB Design Choices**

Regression trees with Sharpe ratios as dependent variable and design choices as explanatory variables. Monthly data for the full sample period from 1963:07 to 2020:12. Boxplots around the end nodes describe the distribution of Sharpe ratios for a particular path along the regression tree. The number on top of each plot counts the number of occurrences along a given path.



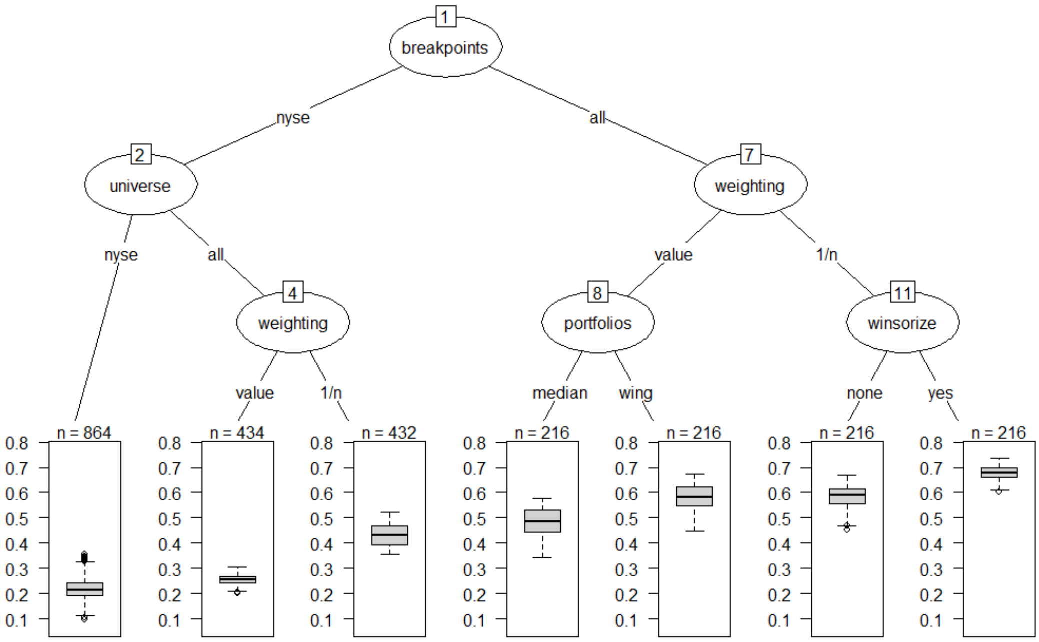
when building long/short factor portfolios in asset pricing research.

The results of our regression trees for SMB and HML are found in Figures 3.3 and 3.4. We find that Size-factor returns offer the best risk-adjusted performance when not being winsorized and using NYSE-only breakpoints and universe; which are - with the exception of the NYSE-only universe - exactly Fama and French's choices. In fact, the SMB regression tree identifies the universe split as an important design parameter in different parts of the tree and the decision to use NYSE-only breakpoints but include NASDAQ and AMEX stocks for the original model seems to be motivated by the better results for the HML factor, as NYSE-listed stocks tend to be substantially larger than in the rest of the universe. This also explains the negative correlation between HML and SMB variations. A large universe is ceteris paribus good for HML but less so for SMB. Fama and French themselves discussed inconclusive SMB results for small buckets in their 5x5 test portfolios.

For HML, our results indicate it performs best when built with the full-universe break-

Figure 3.4: **HML Design Choices**

Regression trees with Sharpe ratios as dependent variable and design choices as explanatory variables. Monthly data for the full sample period from 1963:07 to 2020:12. Boxplots around the end nodes describe the distribution of Sharpe ratios for a particular path along the regression tree. The number on top of each plot counts the number of occurrences along a given path.



points (rather than the standard NYSE-only breakpoints). In contrast to SMB, better performance for HML is achieved with a larger universe, equal-weighting and winsorization. This confirms the common wisdom that the value effect can suffer from outsized returns of a few small but heavy-weight stocks. The R^2 from a regression tree with continuous dependent variables can still be calculated as the squared correlation coefficient between realized and fitted Sharpe ratios. It amounts to 0.91 for the HML tree and 0.82 for the SMB tree. Our regression trees capture the cross sectional variation in risk-adjusted performance well.

After reviewing single factors we now turn to a combination of such factors in factor models.

3.5 Investing into Factor Models

The dispersion of single factors Sharpe ratios shown in the previous section suggests a significant performance impact of design and implementation choices. Will this carry over to factor models or do design choices diversify away once we construct a combined factor portfolio? After all we have shown in the previous section that Sharpe ratios for HML and SMB display negative correlation. Excellent choices for one factor turned out to be poor for the second factor. This section will take an investors view, i.e we look at in and out-of-sample Sharpe ratio variations of factor models and their statistical significance.

To reflect the investors perspective we use factor model Sharpe ratios as our performance metric in the spirit of Barillas and Shanken (2017). The in-sample optimal weight (w^*_{is}) for the self financing long/short factors can be written in the canonical form for unconstrained portfolio optimization

$$w^*_{is} \propto \Sigma_{is}^{-1} \mu_{is} \tag{3.1}$$

where μ_{is}denotes the 3×1 vector of in-sample average excess returns for all three factors and Σ_{is}describes the corresponding variance covariance matrix. From this the in-sample Sharpe ratio follows as

$$SR_{is} = \frac{\left(\Sigma_{is}^{-1}\mu_{is}\right)^T \mu_{is}}{\left(\left(\Sigma_{is}^{-1}\mu_{is}\right)^T \Sigma_{is} \left(\Sigma_{is}^{-1}\mu_{is}\right)\right)^{1/2}} \tag{3.2}$$

This is typically simplified as

$$SR_{is} = \left(\mu_{is}^T \Sigma_{is}^{-1} \mu_{is}\right)^{1/2} \tag{3.3}$$

Out-of-sample optimal weights (w_{oos}^*) are set to equal in-sample optimal weights in order to avoid look ahead bias.

$$w_{oos}^* = w_{is}^* \tag{3.4}$$

Substituting this into the definition of Sharpe ratio (assuming that μ_{oos} denotes the 3×1 vector of out-of sample average excess returns for all three factors and Σ_{oos}describes the corresponding variance covariance matrix), we arrive at

$$SR_{oos} = \frac{\left(\Sigma_{is}^{-1}\mu_{is}\right)^T \mu_{oos}}{\left(\left(\Sigma_{is}^{-1}\mu_{is}\right)^T \Sigma_{oos} \left(\Sigma_{is}^{-1}\mu_{is}\right)\right)^{1/2}} \tag{3.5}$$

which cannot be further simplified. Without enforcing $w_{oos}^* = w_{is}^*$ we would arrive at unrealistically high out-of-sample Sharpe ratios, as negative returning factors would simply be shorted with hindsight, i.e.,

$$SR_{oos} \neq \left(\mu_{oos}^T \Sigma_{oos}^{-1} \mu_{oos}\right)^{1/2} \tag{3.6}$$

As in the single factor case we investigate in Table 3.5, whether alternative design choices lead to superior excess returns per unit of risk. In-sample, *the choice* provides close to median performance with a Sharpe ratio of 0.76. Design choices can lead to Sharpe ratios as low as 0.48 and as high as 1.25. Fama and French could have (in-sample) reached much higher Sharpe ratios at the time, which provides evidence that their objective was not to maximize risk-adjusted returns. Conditional out-of-sample performance declines across all performance bucket percentiles. It is however hardly a statistically significant drop in Sharpe ratio as indicated by our bootstrapping statistics. A corollary of this result is the high degree of persistence across in and out-of-sample results. The best in-sample design variations are also likely to end up among the best out-of-sample specifications, such that few specifications see a large enough drop in performance to reach significant levels.

Factor model Sharpe ratios might be interesting on its own. However, what matters most to investors that depart from the single-beta to a multi-beta world is the performance of an investment into the three-factor models compared with an investment into a broad equity market portfolio. This is also relevant from an asset pricing perspective, as the market portfolio should ex-ante plot below the efficient frontier spanned by market, size and value factor. Ex post, a multifactor portfolio should be closer to the tangency portfolio and display higher rewards per unit of risk than the market portfolio.[16] For this purpose, we test whether the out-of-sample Sharpe ratios for our factor model variations (multifactor-world) exceed the Sharpe ratio from investing into a passive market portfolio (single factor-world). For each design choice, we compare the out-of-sample Sharpe ratio of the Three-Factor model with the out-of-sample Sharpe ratio of the Market Factor. The statistical significance (t-value) is derived from the test on Sharpe ratio differences by Ledoit and Wolf (2008). We employ their non-parametric (iid) bootstrap with 5,000 resamplings and plot the results versus in-sample Sharpe ratios for each design choice in Figure 3.5. Larger in-sample factor model Sharpe ratios lead to higher out-of-sample t-values for a test on Sharpe ratio difference between Three-Factor model and Market factor. While 40% of all variations outperform the market portfolio (t-value for Sharpe ratio difference is positive), this outperformance isn't statistically significant for all models during the out-of-sample period (1992:1 to 2020:12). This is partially caused by low correlations between market portfolio and factor portfolios and the high volatility of factor portfolios which makes Sharpe ratio differences less significant.

Finally, we also find that higher in-sample Sharpe ratios result (on average) in larger out-

[16] See Cochrane (1999) for a review. The average investor (holding the market portfolio by definition) is willing to give up risk-adjusted returns as he cannot take the risks associated with asset pricing factors. These type of risks are negatively correlated with the marginal utility from consumption. Plotting average returns versus risks is insensitive to the investors ability to tolerate the covariances with investors wealth.

of-sample t-values for the Sharpe ratio difference between the market and factor portfolios. Data mining is somewhat successful and "better" in-sample variations also tend to display superior out-of-sample performance. This effect seems to plateau for variations in the highest in-sample Sharpe ratio decile (above 1.1), which could be due to overfitting (in this region).

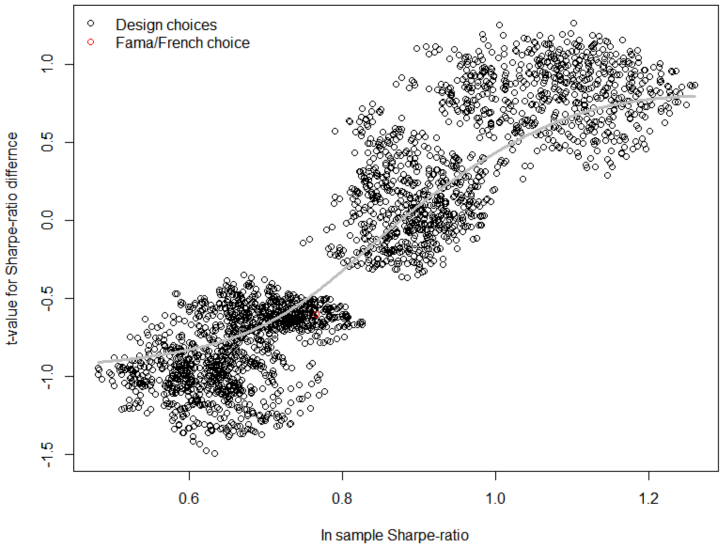
Table 3.5: **Design Choice and Factor Model Performance (Sharpe Ratio)**

We calculate in-sample factor model Sharpe ratios (in-sample tangency portfolio weights times in-sample performance) versus out-of-sample Sharpe ratios (in-sample tangency portfolio weights times out-of-sample performance). Out-of-sample Sharpe ratio are sorted conditional on their in-sample Sharpe ratio. For each design choice we bootstrap the 95%, 97.5% and 99% upper-bound Sharpe ratio consistent with out-of-sample factor model return variations and compare it with its corresponding in-sample ratio. If in-sample Sharpe ratios are higher than the upper confidence bound for resampled pricing errors, we count it as significant.

	1963:7 to 1991:12			1992:01 to 2020:12			% of model design choices with significant Sharpe ratio decay		
Percentile	1st perc	50th perc	99th perc	1st perc	50th perc	99th perc	95% conf.	97.5% conf.	99% conf.
(worst) 0-10	0.484	0.572	0.597	0.329	0.398	0.475	0.000	0.000	0.000
10-20	0.598	0.622	0.643	0.304	0.391	0.473	0.772	0.000	0.000
20-30	0.643	0.658	0.676	0.313	0.413	0.491	7.336	0.000	0.000
30-40	0.677	0.702	0.724	0.316	0.428	0.488	20.000	4.615	0.000
40-50	0.725	0.755	0.801	0.332	0.434	0.610	53.668	15.058	0.000
50-60	0.804	0.841	0.871	0.410	0.600	0.764	24.710	8.108	0.000
60-70	0.872	0.899	0.929	0.489	0.589	0.871	48.846	16.154	0.769
70-80	0.930	0.961	0.994	0.539	0.668	0.873	40.541	15.830	2.703
80-90	0.995	1.060	1.105	0.658	0.807	0.901	14.286	1.158	0.000
(best) 90-100	1.107	1.151	1.252	0.674	0.804	0.892	66.538	34.231	7.308
Fama and French		0.780			0.573		yes	no	no

Figure 3.5: **Outperformance versus Market Portfolio**

We plot t-values on Sharpe ratio differences (between out-of-sample factor model performance and market portfolio). The test procedure follows Ledoit and Wolf (2008) using a non-parametric iid bootstrap with 5,000 resamplings.



3.6 Design Choices and Their Impact on Asset Pricing

This section will follow up with an asset pricing view. The important questions for this sections are: Do we find substantial variation in the asset pricing ability of alternative factor model design choices? Is this variation exploitable, i.e., does it allow us to build better factor models? How well does the FF3F model perform in an asset pricing context? Is its ability to price assets dependent on design choices or is the implementation of minor importance?

To answer these questions we test the ability of each factor model to price the 25 Size-Value test portfolios provided on Ken French's website in- and out-of-sample. Our analysis follows the same layout as in the single factor case. We define the model pricing error as the quadratic form of risk-adjusted excess returns. Following the work by Gibbons et al. (1989) we combine the vector of intercepts from regressing test portfolio excess returns on factor returns originating from design choices (α), and the residual covariance matrix (covariance matrix of regression residuals for all 25 regressions, Ω) into a single mispricing measure:[17]

[17] It is important to notice that a given set of design choices applied symmetrically to all factors. If a particular design choice entails equal weighting across all NYSE stocks, the market factor as well

$$GRS_{is} = \alpha_{is}^T \Omega_{is}^{-1} \alpha_{is} \tag{3.7}$$

Economically this equals the squared information ratio of a factor-neutral long/short portfolio that takes positions in individual test portfolios with perfect knowledge of average excess returns and their covariance structure. In accordance with the literature, we calculate pricing error with a look ahead bias, i.e., we always estimate factor exposures using the same sample information that we use for calculating pricing errors and their covariance structure such that

$$GRS_{oos} = \alpha_{oos}^T \Omega_{oos}^{-1} \alpha_{oos} \tag{3.8}$$

In Table 3.6, we repeat or previously introduced format and sort factor models (derived from specific design choices) into buckets based on their in-sample pricing capabilities (here in terms of the GRS pricing error). We then compare the results with each bucket constituents out-of-sample pricing error. First,we find that *the choice* performs very well in-sample as it outperforms 72.3% of all alternative design choices. This contrasts with our results using in-sample Sharpe ratio as performance metric where *the choice* is close to the median. When making *the choice,* Fama and French (1993) obviously have been more focused on asset pricing rather than asset management. Investment managers should take a note. However, similar to the case of individual factors, out-of-sample results deteriorate considerably. All design variations display similar and large pricing errors. *The choice* deteriorates in ranking versus its peers. We do not find persistence, i.e., lower in-sample pricing error do not lead to lower out-of-sample pricing error. Virtually all choices deteriorate in terms of pricing error and *the choice* sits robustly in the middle.[18] Out-of-sample, design choices do not play an important role. Instead, the impact of design choices is likely to be a second order effect when compared to the static nature of the three factor model which does not accommodate new upcoming factors.

as HML and SMB will employ equally-weighted NYSE stocks.

[18] See also the work by Stambaugh and Yuan (2017); Fama and French (2015); Hou et al. (2015).

Table 3.6: **Out-of-Sample Increase of Mispricing in Factor Models**

Three-factor models related to a particular design choice are sorted into buckets (from worst to best) conditional on their in-sample pricing error. Percentiles for in-sample and out-of-sample pricing errors for each of these buckets are given. For each design choice, we then bootstrap the 95%, 97.5%, and 99% lower-bound pricing error (smaller is better). If a design choice displays an even smaller in-sample pricing error it is deemed significantly lower. Pricing error is defined as the quadratic form of a vector of estimated test portfolio alphas and the variance-covariance matrix of residuals from the corresponding multifactor regression.

	1963:7 to 1991:12			1992:01 to 2020:12			% of model design choices with significant pricing error increase		
Percentile	1st perc	50th perc	99th perc	1st perc	50th perc	99th perc	95% conf.	97.5% conf.	99% conf.
(best) 0-10	0.071	0.099	0.107	0.319	0.356	0.420	100.000	0.000	0.000
10-20	0.107	0.114	0.119	0.324	0.357	0.474	100.000	0.000	0.000
20-30	0.119	0.122	0.124	0.314	0.354	0.430	100.000	0.000	0.000
30-40	0.125	0.127	0.129	0.308	0.353	0.428	100.000	0.000	0.000
40-50	0.129	0.132	0.134	0.309	0.351	0.421	100.000	0.000	0.000
50-60	0.135	0.138	0.142	0.308	0.348	0.419	100.000	0.000	0.000
60-70	0.142	0.146	0.150	0.311	0.352	0.423	100.000	0.000	0.000
70-80	0.150	0.154	0.161	0.310	0.351	0.426	100.000	0.000	0.000
80-90	0.161	0.173	0.186	0.317	0.363	0.404	100.000	0.000	0.000
(worst) 90-100	0.187	0.205	0.255	0.317	0.357	0.386	78.846	0.000	0.000
Fama and French		0.123			0.348		yes	no	no

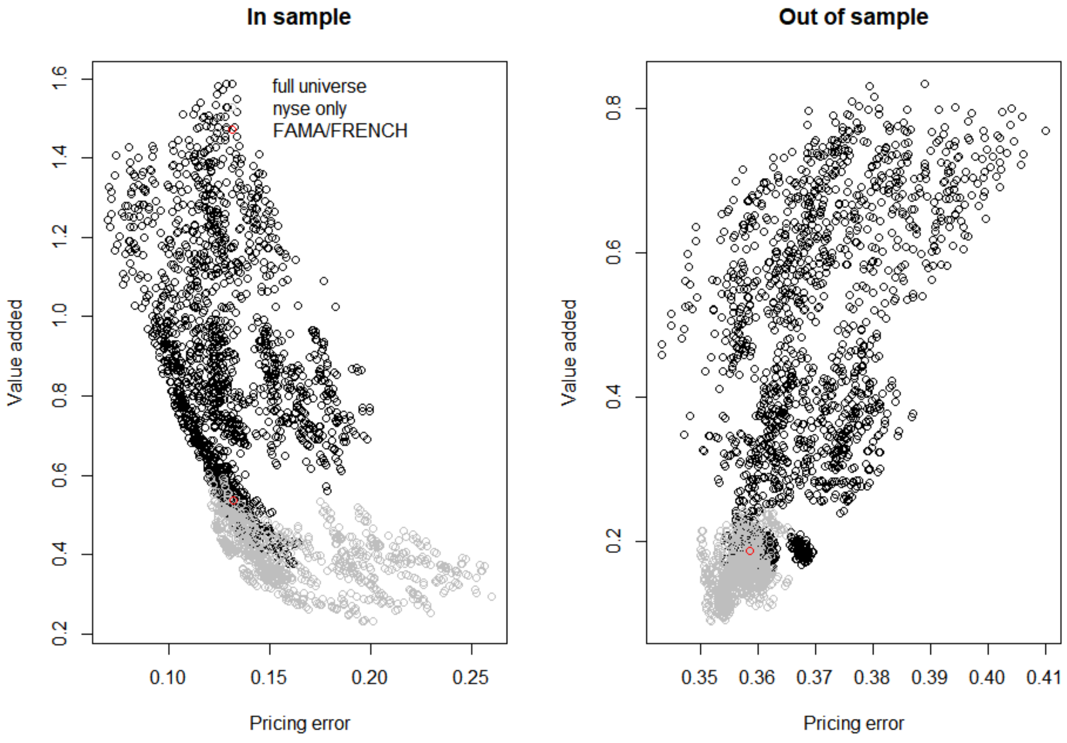
How significant are these results? We apply the same bootstrapping approach as in the previous section to test for significance of out-of-sample pricing error increase.[19] Results indicate that out-of-sample pricing errors are significantly higher as long as we do not impose higher levels of confidence. This did not remain unnoticed within academia as the development of new factor models can tell. Out-of-sample pricing errors for the best in-sample decile are even slightly higher in the extremes than for buckets containing worse in-sample variations.

Recent asset pricing papers loosen the dichotomy between asset pricing and asset management applications. While a factor model consists of more than one factor, we can summarize it with a single return stream, i.e., the return stream arising from the tangency portfolio. This portfolio is both useful for asset pricing and asset management. The beta pricing theorem by Roll (1977) specifies, that by finding a portfolio on the true efficient frontier, we also find an asset pricing model (pricing kernel). The tangency portfolio (maximum Sharpe ratio portfolio) will correctly price the asset universe a given efficient frontier is derived from. Within asset management applications, the tangency portfolio is of large interest as it quantifies how much an investor would benefit by holding a particular factor combination, i.e., design choice. Which perspective was more important for Fama and French when they made *the choice*? For this purpose we look at the relationship of in-sample Value-added (which is defined as squared Sharpe ratio) versus pricing error for all design choices. Figure 3.6 shows the results both in- and out-of-sample. Each circle represents a particular design choice. In addition we marked circles dependent whether the full universe or only NYSE stocks have been used to construct factor returns. In-sample, we find that the *the choice* (red circle) displays a comparatively small pricing error at the expense of lower value added when compared to other design choices. Unsurprisingly, Sharpe ratio maximization seems to be an unlikely motive for *the choice*. Notably, it becomes clear that only variants using the full universe of NYSE, NASDAQ and AMEX reach an in-sample Value-added of above 0.6. For the full universe, higher Sharpe ratios typically come with lower pricing errors. There was

[19] To test whether in-sample pricing errors are significantly lower than out-of-sample errors, we repeatedly resample both test portfolios and factor returns, rerun the pricing equations and calculate a new resampled version of equation 3.8.

Figure 3.6: **Value-added versus Pricing Error**

Value-added (squared Sharpe ratio) versus pricing error for in- and out-of-sample periods. Different universes are marked black (all stocks) or gray (NYSE stocks only)



room for optimization, which was resisted by Fama and French (1993). Out-of-sample this picture falls apart. The range of obtainable pricing errors shrinks dramatically rendering all specifications across both universe decisions less effective out-of-sample. Both in- and out-of-sample we find that using the full universe improves Value-added and is associated with allowing the lowest achievable pricing errors. Using the smaller universe leads out-of-sample to smaller variations in Value-added and pricing error as design choices seem to have limited impact on the smaller universe.

What is a plausible interpretation for our results? Let us assume that the true factor pricing model for our test portfolios contains three additional, new factors, while the old factors become at the same time less important (i.e., displaying lower betas and lower average returns). We would expect that design choices become meaningless with regards to pricing errors but maintain their significant influence on Sharpe ratios. Mispricing is out-of-sample

Table 3.7: **Pricing Error using Principal Components**

We extract the principal components from the correlation matrices of design choices for the 2,592 HML and SMB implementations, respectively. We calculate in-sample and out-of-sample pricing errors using the first one to five principal components. Pricing errors are calculated using the same amount of principal components for HML and SMB. Bootstraping with 5,000 repetitions is applied to test for significant improvements of the principal component pricing models versus the FF3F model. We give the percentage of bootstrapped iterations with lower pricing error than the FF3F model.

	1963:7 to 1991:12		1992:01 to 2020:12	
#Principal Components	Pricing Error	% of bootstrapped model design choices with pricing error decrease to FF3F	Pricing Error	% of bootstrapped model design choices with pricing error decrease to FF3F
1	0.103	96.4%	0.356	37.0%
2	0.084	85.4%	0.300	69.0%
3	0.077	88.0%	0.244	93.1%
4	0.091	65.7%	0.221	96.8%
5	0.098	59.6%	0.236	90.8%
FF3F	0.117	—	0.326	—

dominated by missing factors, not design choices.

This far we have shown how design choices can impact the ability of a factor model to price asset. The observed dispersion in pricing ability suggests that design choices may capture relevant - but different - aspects of the underlying, unknown pricing kernel. Thus, we close this section asking if the pricing error can be reduced by combining the information content of multiple alternative design choices in one pricing model. Principal Component Analysis is particularly helpful in this context as it allows us to derive a reduced-form pricing model out of the variety of design choices. For this purpose, we use the principal components of the 2,592 HML and SMB factor implementation choices[20] in an alternative pricing model and compare its pricing error to the FF3F model (compare Table 3.7). Extracting the first

[20] The first principal component of the HML implementations captures 88.4% of the variation in returns across the design choices increasing to 96.4% for the top five principal components. Similar figures are found for the SMB factor. This result indicates that the substantial dispersion in performance across the design choices can be captured with few orthogonal implementations.

principal component for each universe of HML and SMB variations, the in-sample pricing error of 0.103 is smaller than the respective number from the FF3F model (0.117). We interpret this as further evidence, that *the choice* is highly efficient in spanning the space of alternative model specifications but we do acknowledge that in-sample a PCA-based approach delivers marginally stronger results. Adding further principal components does improve the results versus the base model further up until the point three PCAs are used. However, there is a significant instability in these results related to the in-sample fitting which is revealed by a decreasing significance in our bootstrapping approach compared to the one PCA case. Out-of-sample[21], the pricing error using the first principal component is 0.356 and also in line with *the choice.* The addition of further principal components, however, can lead to significant out-of-sample improvements. Using the first three principal components for HML and SMB, respectively, results in a pricing error of 0.244 which is significantly lower (at the 93% confidence level) than for *the choice.* Thus, similar to our results for factor performance, we find that design fishing has some merit out-of-sample when compared to *the choice.* However, on an absolute basis pricing errors still deteriorate compared to in-sample.[22]

3.7 Summary

We investigate the importance of implementation (design) choices for asset pricing factors using the Fama and French (1993) three-factor model (FF3F). Our usage of the FF3F as reference model is motivated by its long history of out-of-sample data (relative to more recent factor model alternatives) and its pivotal importance for the empirical asset pricing literature. This paper arrives at three main conclusions.

[21] Out-of-sample principal component performance is calculated using in-sample eigenvectors on out-of-sample design choice factor returns.

[22] While PCA is our preferred choice to extract the information content captured in the cross-section of design choices, we can also apply other methodologies. One alternative approach to try improving on the FF3F model would be to pick among the 2,592 design choices for HML and SMB, respectively, the combination that performs particularly well when used for asset pricing purposes. In other words, we do not enforce anymore a consistent portfolio construction for the two factors. Among the 6,718,464 possible permutations between independent SMB and HML implementations we find 32.2% have a lower (i.e., better) pricing error than the FF3F model. However, in the group of those variations that are better in-sample, only 25.0% also show a better pricing ability out-of-sample, giving evidence for the instability of this alternative approach.

First, we document the substantial impact of design choices on the dispersion and persistence of Sharpe ratios for both individual factors as well as combined factor models. Our results tie in nicely with Harvey et al. (2016) and Hou et al. (2020) who illustrate the impact of datamining on significance levels in factor research. Substantial dispersion reminds us that design choices consume degrees of freedom for testing multiple statistical hypotheses. This observation highlights that factor models with different design choices are somewhat difficult to compare. It is hard to establish if differences are driven by economics alone or features driven by factor and design fishing. Performance persistence meanwhile supports another conclusion by Harvey et al. (2016) who state that "the optimal amount of data mining is not zero". Our results agree with this perspective. Additionally, we find evidence that design choice-induced dispersion does not affect factors equally or randomly. In the case of the FF3F, Sharpe ratio changes across design choices are negatively correlated for SMB and HML. This has two important implications: Observed dispersion is lower on model level than on single factor level and design choices need to be considered when evaluating additional factors or anomalies in combination with the FF3F.

Second, we find that the original specification of Fama and French has been derived from the perspective of asset pricing research as it does not appear to be optimized to deliver better risk-adjusted returns. *The choice* by Fama and French offers superior in-sample pricing ability, outperforming 72.3% of alternative design choices. At the same time it only provides an average in-sample Sharpe ratio. Out-of-sample the decay in the three factor models ability to price the cross section of test portfolios affects all design choices equally. *The choice* made by Fama and French displays no out-of-sample bias, but rather shows that missing factors can not be compensated with clever design choices.

Finally, we find that only 40% of our variations offer higher out-of-sample risk-adjusted returns than an investment into the market portfolio. However, given the variability of factor returns and the low correlation of the FF3F model returns with market portfolio returns, this evidence comes with weak statistical significance and warrants further attention.

If he's still in, I'm still in!

How Reddit posts affect GameStop retail trading

André Betzer [*] Jan Philipp Harries [*] [†]

May 2021

Abstract

In January 2021, the stock price of NASDAQ-listed GameStop Corporation surged more than thirty-fold following frenzied discussions on a Reddit forum. While Social Media-organized retail trading isn't a new phenomenon, the magnitude of the resulting swings in GameStop's share price in combination with a short- and/or gamma squeeze scenario is unprecedented. Using financial data as well as an extensive dataset of Reddit posts, we show that Reddit posts lead to increased retail trading activity in GameStop shares and introduce a new option-based measure for retail trading proportion which could help academics, regulators and investors to track and analyze similar developments in the future.

Keywords: GameStop, Retail Trading, Sentiment Trading, Market Structure

JEL Codes: D91, G14, G41

[*] University of Wuppertal, Gaußstraße 20, 42119 Wuppertal, Germany

[†] Corresponding Author: harries@wiwi.uni-wuppertal.de

4.1 Introduction

> ***GameStop: How a video game chain was dragged into the war on Wall Street****. An army of retail investors pushed a struggling business to a $28bn market capitalisation.*
>
> Financial Times Headline, January 28th, 2021

The literature on retail trading in financial markets has already investigated some important insights into the characteristics of retail trading and the predictability of retail investor trading on future stock returns (e.g. Kaniel et al., 2008; Han and Kumar, 2013; Boehmer et al., 2021). In this context of retail trading, the recent situation around GameStop is uniquely suited to answer the so far unanswered research question whether activity on a social media platform can have a direct impact on the retail trading proportion (hereafter RTP) of a firm's common stock and its derivatives.

Starting in 2020, users of the message board r/WallStreetBets[1] (WSB) on the social media platform Reddit turned their eye on the stock of struggling video game retail company GameStop. While only a few users discussed the stock at first[2], hundreds and thousands of retail investors joined them in early 2021, when the GameStop stock surged due to the expectation of an imminent short squeeze. While GameStop opened in 2021 on January 4th with a price of $19.03, the closing price on January 21th was already $43.03, an increase of more than 100% without any new fundamental information released by the company in the meantime. However, the real surge had barely started by then: In the following five trading days, the share price increased 10-fold again and reached a top of $483 in the morning of Thursday, January 28th, before major brokers disabled market participants ability to open new or increase existing positions in GameStop. Huge losses of GameStop-shorting hedge funds, margin calls of unprecedented size and the failure of established risk management models led the CEO of IB, one of the biggest American brokers to the statement that "we

[1] The URL of Reddit message boards, so-called subreddits, is prefixed with "r/" which is why users often refer to subreddits like WSB as "r/WallStreetBets" to avoid disambiguity. See chapter 4.2.3 for more information on Reddit and WSB.

[2] Thereof famously user u\deep[*expletive*]value, who early on invested a large part of his portfolio in GameStop shares, posted regular updates about his investments and was cited to testify in front of a Congressional Hearing about his involvement later on.

have come dangerously close to the collapse of the entire system and the public seems to be completely unaware of that, including Congress and the regulators”[3].

While several other factors, which are partially discussed below, may have contributed to the surge or could have ignited the initial interest in this specific company, media reports and online discussion suggest that Reddit community r/WallStreetBets played a crucial role in this situation, which is also confirmed by our results. The chair of the SEC, Gary Gensler, acknowledged that "this winter’s events also highlighted the rapidly changing face of social media and its intersection with our capital markets"[4]. However, the nature of the relationship between social media posts and trading activity is less clear than it seems and data availability and noisiness problems are complicating the scientific and forensic examination of events.

Due to the manifold implications for our understanding of asset prices and markets, multiple recent working papers try to shed light on different aspects of the GameStop surge: Long et al. (2021) classify Reddit posts into sentiment categories and find that "both tone and number of comments influence GameStop intraday returns". Vasileiou et al. (2021) find that "the skyrocketing performance of GameStop shares causes the increased interest for the GameStop short squeeze" using Google searches and intraday data while Umar et al. (2021) use Twitter data to study sentiment driven pricing. Others investigated the effect of trading restrictions, try to classify participating investors, extend behavioral models or analyze possible regulatory reactions (e.g. Jones et al., 2021; Hasso et al., 2021; Pedersen, 2021; Angel, 2021).

In contrast, we limit our analysis to the specific and so far unanswered question whether Reddit posts (and thus Social Media activity) are indeed a driver of retail trading activity. Regardless the direction and exploitability of trading and pricing changes, the existence of a direct link between social media activity (without the dissemination of material new information) and stock buying from a specific group of investors at valuation levels which far exceed any reasonable fair value is at odds with conventional economic theory and the

[3] see `https://www.cnbc.com/2021/02/17/interactive-brokers-chairman-thomas-peterffy-on-gamestop-frenzy.html`.

[4] According to his written testimony before the US House Commitee on Financial Services on May 6th, 2021, which can be found at `https://financialservices.house.gov/uploadedfiles/hhrg-117-ba00-wstate-genslerg-20210506.pdf`.

efficient market hypothesis. While prior literature already showed that stocks with a high proportion of retail traders often exhibit lottery features and "attract retail investors with strong gambling propensity" (Han and Kumar, 2013), we are, to the best of our knowledge, the first to show empirically how activity on a social media platform affects the retail trading proportion over time, using a dataset of more than 40 million Reddit posts and employing multiple proxies for retail trading activity by using stock and option tick data for GameStop from January 2020 until March 2021.

Thus, our contribution is twofold: First, we show that an increase in Reddit posts on GameStop is followed by an increase in the ratio of retail trading proportion of GameStop. An increase of 50% in WSB comments on GameStop will lead to an increase of the retail trading proportion of approximately 0.7% in the following 30-minute window. While economically small, the effect is larger than what was found in comparable studies and robust and consistent over multiple retail trading classifications.

Second, we compare multiple procedures to identify retail trades in a highly relevant real-world case study. Besides measures that are based on odd-lots, trade size and subpenny improvements, we introduce a new option-based measure which relies on one-contract trades and hasn't been used in this context before.

The remainder of this paper is organised as follows. Section 4.2 specifies the research hypotheses and contains a brief literature review on Retail Trading. Section 4.3 explains data and methodology, followed by the discussion of empirical results in Section 4.4. Section 4.5 provides some directions for future research and concludes with a summary.

4.2 Retail Trading and r/WallStreetBets

4.2.1 The importance of Retail Trading

Han and Kumar (2013) find empirical evidence that retail investors in contrast to institutional investors prefer "stocks with high volatility, high skewness and low prices". In addition, the authors document that the characteristics of retail traders are similar to the characteristics of investors who prefer lottery stocks (those are often younger and male and have a lower income and lower education compared to other investors), as has been shown in Kumar and Lee (2006).

The empirical evidence on the predictability of retail investor trading on future stock returns is mixed. While many early studies in this strand of literature such as Barber et al. (2006) find that individual investors trading provide no information for equity markets and prices, more recent studies such as Barrot et al. (2016) or Boehmer et al. (2021) find that retail investor trading can predict the cross section of future stock returns. One important reason for these different findings in the literature can be the identification strategy of retail trades vs. institutional trades and hence the potential misclassification of such trades.

While many existing studies use trade-size as proxy for retail trading activity, Boehmer et al.'s (2021) main contribution is to provide a more accurate measure for retail trading based on the publicly available TAQ database. They "identify transactions as retail buys if the transaction price is slightly below the round penny, and retail sells if the transaction price is slightly above the "roundpenny", a feature that makes retail trades different from institutional trades. In this paper, we employ Boehmer et al.'s (2021) measure and other proxies for retail trading activity that have been used in the literature so far.

Another interesting aspect of the situation around the GameStop surge and a reason while our contribution is highly relevant is the general increase in retail trading, partially attributable to the COVID-19 pandemic which limited many day-to-day (out-of-home) activities for a big part of 2020 and in early 2021. Analysts from Credit Suisse estimated in February 2021 that retail trading as a share of overall market activity had nearly doubled from between 15% and 18% to over 30% since the start of 2020.[5] Another important reason for sharply increased retail trading has been the decision of many retail-focused US brokers to drop commissions in the fall of 2019. Robinhood, an app-based broker with more than 3 million app downloads in January 2021[6], is the most notable of a new kind of brokers which gamify trades and make stock (and option) trading available to a new demographic which also exhibits high social media affinity and activity. Ozik et al. (2021) confirm that "access

[5] Source: `https://www.cnbc.com/2021/02/13/why-retail-investors-are-here-to-stay.html`.

[6] According to data provider SimilarWeb.

to financial markets facilitated by fintech innovations to trading platforms, along with ample free time, are significant determinants of retail-investor stock-market participation." The importance of this development for market mechanics is underscored by van der Beck and Jaunin's (2021) finding that "despite their negligible market share of 0.2% [...] Robinhood traders account for 10% of the cross-sectional variation in stock returns during the second quarter of 2020.".

4.2.2 Prior Work on the Effect of Sentiment on the Stock Market

A number of researchers have previously examined the role of short-term and long-term sentiment in explaining stock returns and many different proxies have been used (e.g., sentiment based on twitter mentions, news mentions, investor surveys, or analyst recommendations). Daniel et al. (2002) established that psychological biases affect investor behavior and prices. Using this as a starting point, researchers noticed deviations from expected fair values due to changes in investor sentiment: Tetlock (2007), García (2013) and more recently e.g. Kumar et al. (2020) found that newspaper sentiment can lead to pricing anomalies. Others, like Baker and Wurgler (2007) and Cornell and Damodaran (2014) (with a case study on Tesla, which happens to also be a very popular stock on WSB in 2020 and 2021) focus on deviations from fundamental value. With regards to social media, the influence of Twitter posts on the stock market has been well researched (see e.g. Behrendt and Schmidt, 2018; Broadstock and Zhang, 2019; Nisar and Yeung, 2018). In chapter 4.2.3 we will also refer to some previous studies about WSB and GameStop.

However, published results about the impact of investor sentiment from various sources are not very consistent. While most published studies claim a significant impact of investor sentiment on prices, others paint a mixed picture (e.g. García, 2013; Ahmad et al., 2016) or find no significant or predictable effect (e.g. Campbell et al., 2012; Behrendt and Schmidt, 2018; Nisar and Yeung, 2018). This is one reason for our conservative approach of focusing solely on the retail trading proportion instead of the effect of WSB posts on stock returns or volatility.

4.2.3 Reddit and r/WallStreetBets

Reddit is a social media platform which was founded in 2005. Like other social media platforms, contributors are able to post content which can then be be commented on by other users. Reddit is a collection of forums, which are called subreddits and where each subreddit is dedicated to a specific topic. WSB, which is now one of Reddit's largest subreddits with more than 10 million subscribers, was created in 2012 and focuses on speculative equity trading.[7] As speculative trading and "gambling" is emphasized, it is reasonable to assume that retail trading activity originating from WSB may exhibit different characteristics than retail trading activity from more conventional sources of investment advice or discussion.

Due to the enormous growth of the r/WallStreetBets community on Reddit and, more recently, the huge media coverage of the GameStop situation, WSB has also drawn attention from other researchers as the data is uniquely suited to analyze open research questions, especially on Retail- and Sentiment-Trading. Most closely related to our analysis, Long et al. (2021) classify Reddit posts into sentiment categories and find that "both tone and number of comments influence GameStop intraday returns". However, these effects are not very strong and our data suggests that textual sentiment classification into emotion-based categories is very difficult for WSB posts, as these contain a lot of different slang, memes and emoticons which are barely understandable for uninitiated readers (or parsers).[8] Related work (e.g. Behrendt and Schmidt, 2018, for Twitter sentiment) shows that using the raw amount of social media posts or mentions, as we do in our analysis, often yields comparable or even better results compared to calculated sentiment scores, especially if the variable of interest is undirected. Additionally our variable of interest, the RTP, is conceptually more directly related to social media activity than the raw returns investigated by Long et al. (2021) and we concentrate on 30-minute windows instead of 1-minute windows which yields us sufficient data for a longer time horizon.

Vasileiou et al. (2021) also concentrate on explanations for stock price returns, finding

[7] However, despite this focus, Pennystocks are banned from discussion on WSB due to the prevalence of pump&dump schemes.

[8] For example WSB users frequently call themselves "autists" or "apes"; an (incomplete) glossary can be found at `https://www.reddit.com/r/wallstreetbets/comments/l7fr21/basic_guide_to_wallstreetbets_culture_for/`.

that "increased trading volume leads to price increases" in the case of GameStop. Umar et al. (2021) also focus on returns and use Twitter and news data instead of Reddit data. Additionally, they confirm our suspicion that option trading played an important role, too, finding that "the put–call ratio strongly and positively affects the GameStop returns prior to the peak of the GameStop saga".

Finally, Bradley et al. (2021) analyze the market consequences of WSB due diligence (DD) posts, finding that these posts lead to significant abnormal returns. While their main focus is on abnormal returns and event studies, they also find an increase in retail trading activity following these specific DD posts using a model introduced by Farrell et al. (2020) and classifying retail trades as in Boehmer et al. (2021). However, as they concentrate on the publication of specific posts which only represent a negligible share of all WSB posts (e.g. 1734 posts out of more than 20 million WSB posts in 2020 were DD posts) and on total retail trades instead of the RPT, our approach illuminates a different, more direct angle of social media-influenced retail trading.

4.3 Data and Methodology

4.3.1 Data Sources

We collect all WSB posts from the start of 2020 to the end of February 2021, in total more than 40 million, thereof 33.2 million in our trading hours-sample. For posts starting in January 2021 we got the posts directly from Reddit's streaming API and for posts in 2020 we used the unofficial Pushshift API which ingests all Reddit posts (see Baumgartner et al., 2020, for more information on the Pushshift dataset and API). Comments are then sorted into GameStop-related and non-GameStop related comments by the following procedure: 1) A comment is GameStop-related if "GameStop" or its ticker "GME" are mentioned in the comment. 2) If neither GameStop nor another stock symbol is mentioned in a comment, we search iteratively in the parent comment or post for the mention of a stock symbol or company name and classify the post as GameStop-related if the first parent stock symbol or company name mention is related to GameStop. In total, we find that 4 million or about

12% of the comments in our sample period are GameStop-related with the share increasing from below 1% for most of the first half of 2020 to 60% at the peak end of January 2021.

We obtain consolidated stock TAQ trade data for GameStop from IB and consolidated option TAQ data from Ivolatility. In our sample, the stock price of GameStop oscillated from a low of $2.57 on 2020/4/3 to a high of $508.02 on 2021/1/28. There were on average 9.8 million GameStop shares and 115.500 GameStop options (corresponding to 11.55 million shares) traded per day.

Table 4.1: Summary Statistics Raw Data

Table 1 reports summary statistics for our raw data. Consolidated Stock Tick data is obtained from TAQ and consolidated Option Tick data from Ivolatility. 'trd' columns contain statistics for the trade count while 'vol' statistics relate to trading volume. Reddit comments are scraped directly from Reddit and for historic comment data partially from the public Pushshift API. Data is then sorted into 30-minute bins corresponding to 13 30-minute bins for each full trading day in the sample. The sample period is Jan. 2020 - Mar. 2021.

	N	Mean	Std	Q1	Median	Q3	Sum
Stock Ticks (trd)	3666	7172	26 313	712	1230	2458	26 294 370
Stock Ticks (vol)	3666	789 652	2 069 896	138 944	259 968	546 593	2 895 063 527
Option Ticks (trd)	3666	1689	5587	68	183	706	7 368 690
Option Ticks (vol)	3666	8352	22 365	634	1781	5053	34 996 936
All Comments	3639	3081	6392	1612	2045	2586	33 203 672
GameStop Comm.	3639	407	2376	0	1	9	4 076 004

Summary statistics on our raw data can be found in table 4.1. We divide our sample period into 3666 30-minute windows with each full trading day containing 13 of these windows (we follow e.g. Sun et al., 2016; Gao et al., 2018; Farrell et al., 2020, with the use of 30-minute intraday windows). Summary statistics as well as results are provided for *vol* variables, relating to traded volume, as well as *trd* variables which denote the total number of trades independent from each trade's volume.

Due to the unprecedented surge in the share price, as well as volume and WSB comments the development over our sample period is pictured in figure 4.1 using a logarithmic scale on the y-axis.

4.3.2 Proxies for Retail Trading Proportion

As we do not conduct a broad cross-section analysis but a case study focused on a specific stock with a highly unusual and very dynamic trading pattern and price and volume development, a single measure is not sufficient to shed light on the connection between Reddit posts and retail trading activity. Bradley et al. (2021) follow the methodology introduced by Farrell et al. (2020) and use the logarithm of the number of retail trades as LHS in their regression analysis. However as our variable of interest is continuous, the number of retail trades exhibits strong autocorrelation and we can't control for singular events, we concentrate our study on the retail trading proportion (RTP) introduced by Han and Kumar (2013) and also used as an additional measure by Farrell et al. (2020) which is defined as the ratio of the retail trading volume in half-hour windows to the total trading volume in the same window for *vol* variables and the ratio of the number of retail trading transactions to the number of all trading transactions for *trd* variables.

To identify retail trades, we employ 4 different procedures leading (combined with vol/trd) to 8 slightly different RTP measures in total.

1. Our first measure $RTP(OL)$ classifies all odd-lot trades as retail trades:

   $$RTP(OL) = \frac{Trades\ where\ Size\ \%\ 100! = 0}{All\ Trades}$$

   This identification is one of the oldest and most established ones and follows e.g. Dyl and Maberly (1992). However, more recently O'Hara et al. (2014) and others warned that odd-lot trading, while still often used by retail traders, is increasingly caused by high frequency or algorithmic traders. As these kind of traders are less likely in very-high volatility environments like GameStop in our sample period and odd-lot trading is still widely used as a proxy for retail trading, we incorporate $RTP(OL)$ in our analysis.

2. The second measure, $RTP(ST)$, refers to small trades as main criterion. Here we follow e.g. Barber et al. (2006) and Han and Kumar (2013), who use a trade size of \$ 5,000 as cut-off value for a classification as retail trade. Han and Kumar (2013)

confirm that their definition "closely captures the preferences and trading activities of retail investors" by comparing it "with actual retail holdings and trading data from a broker". As we only survey a single stock and the average dollar-denominated trade-size explodes due to the sudden surge, we define this version as

$$RTP(ST) = \frac{Trades\ where\ Size < 264}{All\ Trades},$$

with 263 shares corresponding to a trade value of \$ 5,000 at the GameStop price in the start of 2021. However, similar limitations as for $RTP(OL)$ apply.

3. Our third measure $RTP(OC)$ is option-based and thus avoids many of the shortfalls of $RTP(OL)$ and $RTP(ST)$ as automated trading is less prevalent on the option market, mainly due to lower liquidity and bigger spreads. It is based on the observation that retail traders mostly trade single option contracts (one contract corresponds to 100 shares) while institutional traders who use options e.g. to hedge positions rarely trade single contracts:

   $$RTP(OC) = \frac{Option\ Trades\ where\ Size = 1}{All\ Option\ Trades}$$

   Retail option trading is a novel phenomenon and we are - to our knowledge - the first to introduce this option-based measure. While already Battalio et al. (2004) wrote that they "examine one-contract trades separately to isolate retail orders more confidently" a more recent cross-sectional analysis is still missing in the scientific literature. However, research by brokers, e.g. Goldman Sachs, shows that retail option trading and especially one-contract trading increased sharply since the beginning of 2020 with one-contract trades now accounting for 13% of total option volume and even more for popular stocks.[9]. We can confirm this impression and our results indicate that $RTP(OC)$ identifies retail trades even better than the other measures, however broader cross-sectional research is needed to confirm our GameStop-focused results.

[9] See e.g. https://www.bloomberg.com/news/articles/2020-05-22/options-are-now-all-the-rage-for-bored-day-traders-locked-inside or https://www.barrons.com/articles/how-retail-investors-are-fueling-the-nasdaqs-wild-ride-51599866516.

4. Finally, $RTP(MR)$ identifies marketable retail orders as laid out by Boehmer et al. (2021):

$$RTP(MR) = \frac{FINRA\,Trades\ where\ (Price\ \%\ 0.01)\ *100\ in\ ]0, 0.4]\ or\ [0.6, 1[}{All\ Trades}$$

   Trades are classified as retail trades if the TAQ data indicates that they have been reported through a FINRA-facility and are priced just below a round penny (fraction of a cent between 0.6 and 1) or just above a round penny (fraction of a cent between 0 and 0.4). While this classification captures retail trades reliably due to the regulatory rules around sub-penny price improvements and the increasing internalization of orders by retail brokers it omits limit trades which are not marketable and all trades that are routed to exchanges.

Table 4.2 contains summary statistics about the different RPT measures. The mean proportion of volume identified as retail volume ranges from 9% for $RPT(OC)$ to 35% for $RPT(ST)$ and for the number of transactions from 15% for $RPT(MR)$ to 85% for $RPT(ST)$.

In figure 4.2, one can see that all measures slowly increase in the second half of 2020 and peak end of January 2021, with the increases most pronounced for $RTP(OL)$ and $RTP(ST)$, while $RPT(MR)$ changes least.

For the ratio of the number of retail trades in figure 4.3, $RTP(ST)$ increases significantly less due to the already very high level but the increase for $RPT(MR)$ is more pronounced.

Figures 4.4 and 4.5 show how the different RTP measures relate to each other. For most measures, a significant positive correlation is visible as expected. Only $RPT(ST)$ and $RPT(MR)$ seem to be negatively correlated for the volume ratio. For the number of trades, clusters of 30-minute windows with very high RTP's a visible for different measures. These correspond to the trading days at the start of 2021, where RTP increased strongly compared to previous levels.

Figure 4.1: Logarithmic Chart of GameStop's Stock Price, Traded Volume and Reddit WSB Comments

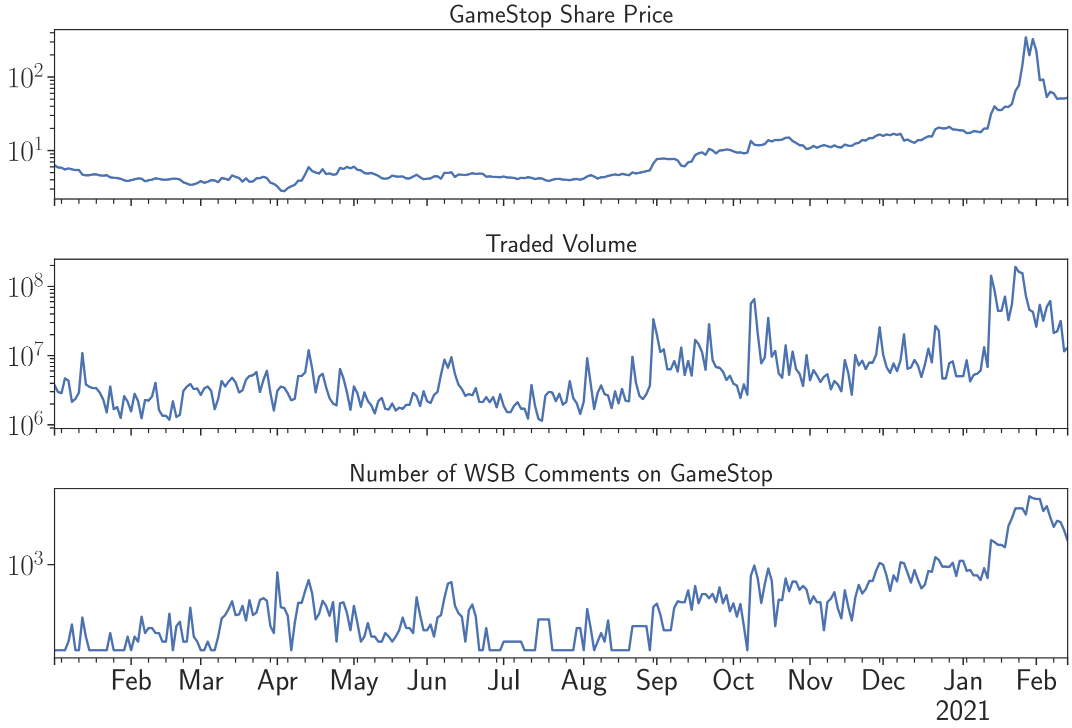


Figure 4.2: Daily Development of RTP, measured as Proportion of Trade Volume (Rolling Average over 5 days)

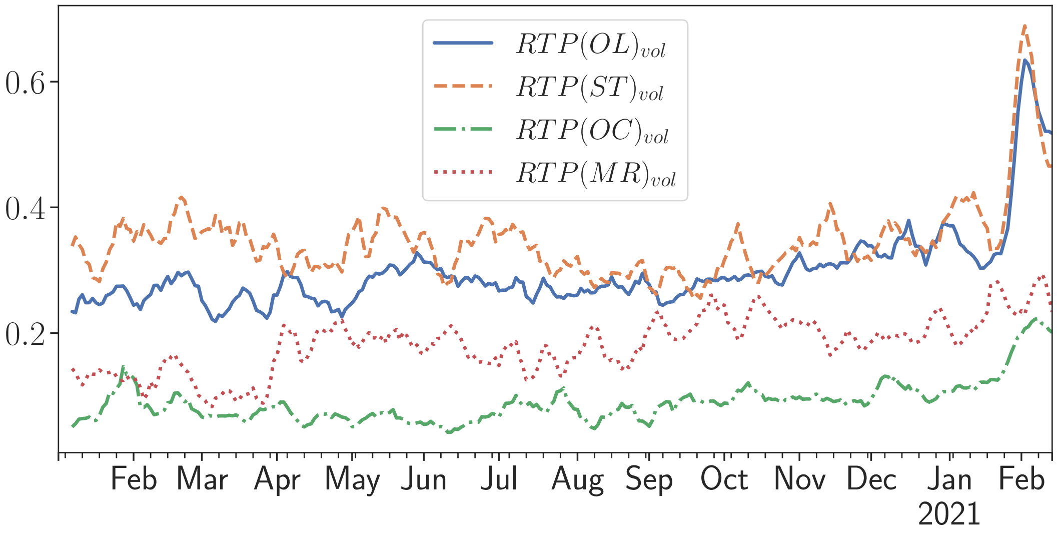


Table 4.2: Summary Statistics RTP Measures

Table 2 reports summary statistics for the four different proxies we employed to measure Retail Trading Proportion, RTP. These are: i) RTP_{OL}, the proportion of oddlot-trades; ii) RTP_{ST}, the proportion of small trades; iii) RTP_{OC}, the proportion of one-contract option trades and iv) RTP_{MR}, the proportion of marketable retail orders as defined by Boehmer et al. (2021). Each observation corresponds to a 30-minute window. Panel I contains statistics for the trade volume while Panel II relates to trading count. The sample period is Jan. 2020 - Mar. 2021.

Panel I: RTP_{vol}						
	N	Mean	Std	Q1	Median	Q3
$RTP(OL)_{vol}$	3651	0.2972	0.1009	0.2318	0.2866	0.3432
$RTP(ST)_{vol}$	3651	0.3474	0.1073	0.2712	0.3344	0.4105
$RTP(OC)_{vol}$	3650	0.0900	0.0727	0.0422	0.0747	0.1182
$RTP(MR)_{vol}$	3651	0.1815	0.0801	0.1224	0.1800	0.2358
Panel II: RTP_{trd}						
$RTP(OL)_{trd}$	3651	0.4992	0.1233	0.4276	0.4858	0.5457
$RTP(ST)_{trd}$	3651	0.8593	0.0478	0.8294	0.8615	0.8879
$RTP(OC)_{trd}$	3650	0.4389	0.1084	0.3704	0.4444	0.5098
$RTP(MR)_{trd}$	3651	0.1459	0.0604	0.1041	0.1410	0.1782

4.4 Results

4.4.1 Regression Analysis of the Effect of WSB Comments on the Retail Trading Proportion

For our first analysis, we explore the predictive effect of Reddit comments on the retail trading proportion using the following regression

$$RTP_t = \alpha + \beta_1 log(1 + RC_{t-1}) + \beta_2 RTP_{t-1} + \beta_3 log(1 + Volume_{t-1}) + \beta_4 R_{t-1} + \epsilon \tag{4.1}$$

with the different variations of RTP dependent variables and the logarithm of the number of Reddit comments in the preceding 30-minute window as main variable of interest. The lagged values of RTP, the logarithm of the trading volume and the absolute return R are used as additional control variables on the right-hand side.

Figure 4.3: Daily Development of RTP, measured as Proportion of Trade Count (Rolling Average over 5 Days)

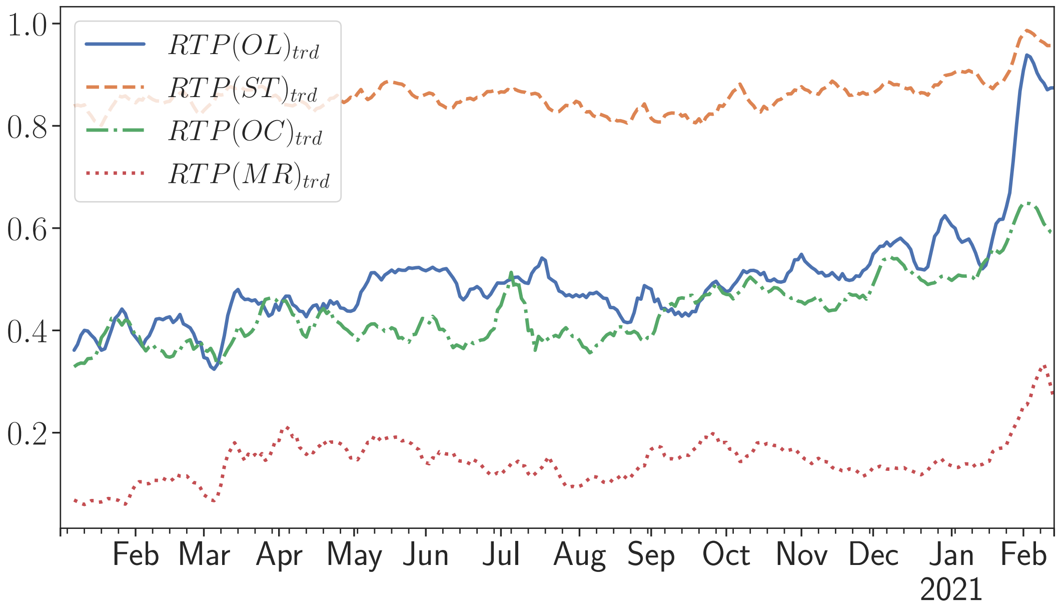


Figure 4.4: Pairplot of RTP Variants, measured as Proportion of Trade Volume in 30-minute Windows

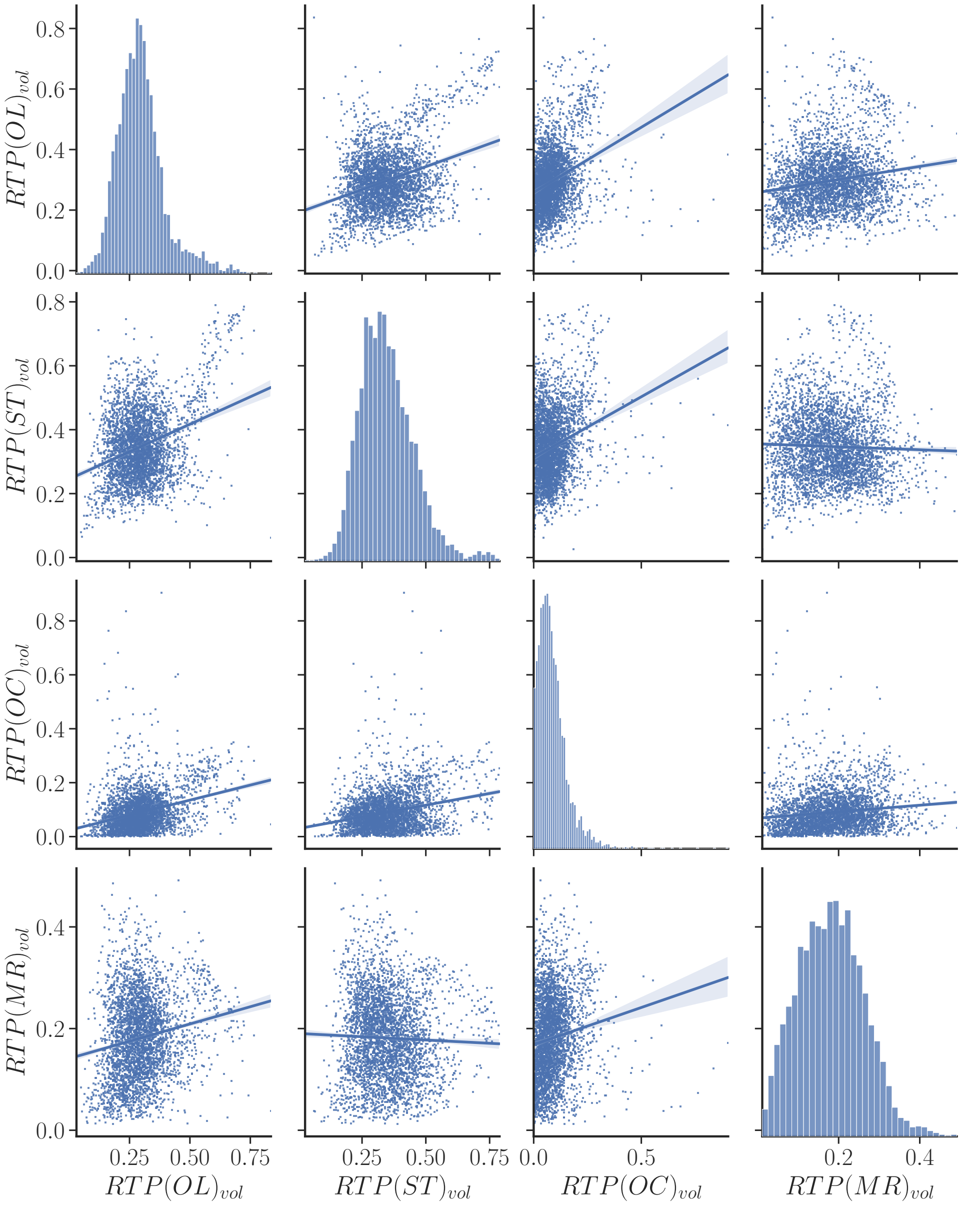


Figure 4.5: Pairplot of RTP Variants, measured as Proportion of Trade Count in 30-minute Windows

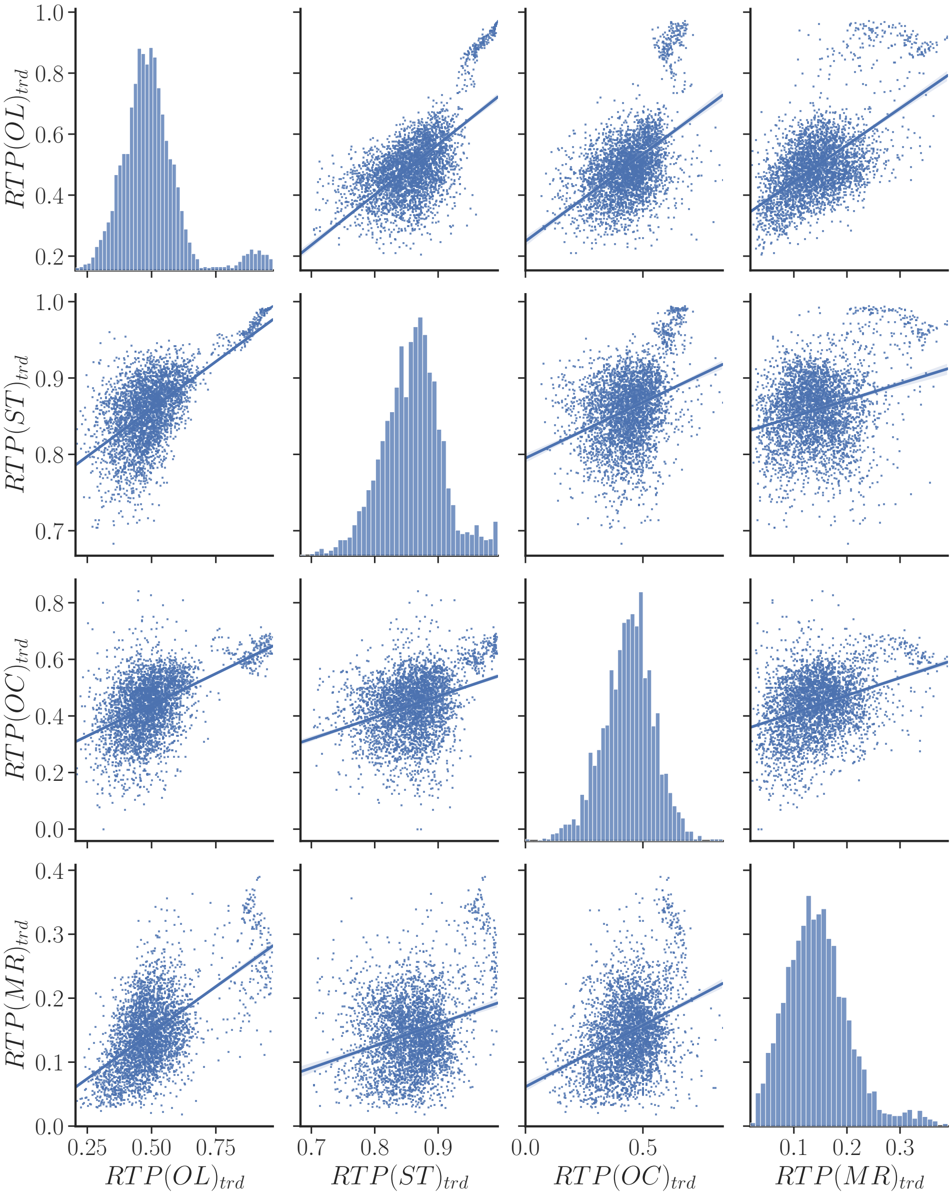


Table 4.3:
How Reddit Comments affect Retail Trading - Regression Estimates I

This table reports the regression estimates, where the dependent variables are different proxies for the Retail Trading Proportion RTP. RTP measures the proportion of retail trading volume to all trading volume in a 30-minute window. The variants are: i) $RTP(OL)$, the proportion of oddlot-trades; ii) $RTP(ST)$, the proportion of small trades; iii) $RTP(OC)$, the proportion of one-contract option trades and iv) $RTP(MR)$, the proportion of marketable retail orders as defined by Boehmer et al. (2021). All independent variables are measured one period ahead of the dependent variable. The independent variable of interest is RC_{t-1}, which denotes the number of all Reddit Comments during the preceding 30-minute window in the WallStreetBets, Stocks and Investing subreddits, where GameStop is explicitly mentioned (either in the post itself or in the parent post). Additionally, the following control variables are employed, each measured for the preceding 30-minute window: i) RTP, a lagged measure of the dependent variable; ii) $Volume$, the trading volume in GameStop shares; iii) R, the absolute return of GameStop shares as a measure for idiosyncratic volatility. Additionally we add Time Fixed Effects (for all 30-minute windows and weekdays) in specifications 2 and 3. We use the Newey and West (1987) methodology with 11 lags to correct standard errors for potential heteroskedasticity and autocorrelation. The sample period is Jan. 2020 - Mar. 2021; independent variables of interest are shown in bold.

Panel I: Proportion of trade volume; no daily independent variables

	$RTP(OL)_t$ x 100		$RTP(ST)_t$ x 100		$RTP(OC)_t$ x 100		$RTP(MR)_t$ x 100	
	Coef.	t-stat	Coef.	t-stat	Coef.	t-stat	Coef.	t-stat
$Intercept$	22.0569***	9.85	43.3142***	13.00	10.3650***	5.47	−8.2661***	−4.59
$log(RC_{t-1})$	**1.7198*****	**16.21**	**1.4540*****	**12.68**	**1.0678*****	**13.43**	**0.2471*****	**3.84**
RTP_{t-1}	32.1279***	16.49	40.6148***	17.76	16.3112***	5.88	38.9150***	21.98
$log(Volume_{t-1})$	−0.3766**	−2.09	−2.0169***	−8.61	−0.3710**	−2.49	1.5037***	10.30
R_{t-1}	8.9859**	2.29	30.7954***	6.42	7.8855***	3.39	−4.7600**	−2.11
Adj. R^2	0.36		0.34		0.16		0.27	
Obs	3556		3556		3554		3556	

$^{*} p < 0.1$, $^{**} p < 0.05$, $^{***} p < 0.01$

Results for the trading volume-based RTP measures can be found in table 4.3. We see a highly significant positive effect of the number of Reddit comments on RPT for all four variations. The comparably small coefficient is expected due to autocorrelation. The result indicates, that e.g. an increase of 50% in WSB Comments on GameStop will lead to an increase of the $RTP(OL)$ of approx. 0.7% in the following 30-minute window. While this doesn't sound huge, the effect is robust and consistent over all RPT definitions. In comparison, Farrell et al. (2020) found an increase of 0.17% in RTP in half-hour windows after the publication of research reports on the news website Seeking Alpha to be highly significant. T-values range from 12.68 to 16.21 for the first three variations and are 3.84 for $RTP(MR)$. Notably, signs for the volume and absolute return controls are swapped for $RTP(MR)$ - while RTP decreases for all other measures after an increase in volume and increases after high absolute returns, we see the opposite effect for $RTP(MR)$.

Table 4.4:
How Reddit Comments affect Retail Trading - Regression Estimates II

This table reports the regression estimates, where the dependent variables are different proxies for the Retail Trading Proportion RTP. RTP measures the proportion of the number of retail transactions to all transactions in a 30-minute window. The variants are: i) $RTP(OL)$, the proportion of oddlot-trades; ii) $RTP(ST)$, the proportion of small trades; iii) $RTP(OC)$, the proportion of one-contract option trades and iv) $RTP(MR)$, the proportion of marketable retail orders as defined by Boehmer et al. (2021). All independent variables are measured one period ahead of the dependent variable. The independent variable of interest is RC_{t-1}, which denotes the number of all Reddit Comments during the preceding 30-minute window in the WallStreetBets, Stocks and Investing subreddits, where GameStop is explicitly mentioned (either in the post itself or in the parent post). Additionally, the following control variables are employed, each measured for the preceding 30-minute window: i) RTP, a lagged measure of the dependent variable; ii) $Volume$, the trading volume in GameStop shares; iii) R, the absolute return of GameStop shares as a measure for idiosyncratic volatility. Additionally we add Time Fixed Effects (for all 30-minute windows and weekdays) in specifications 2 and 3. We use the Newey and West (1987) methodology with 11 lags to correct standard errors for potential heteroskedasticity and autocorrelation. The sample period is Jan. 2020 - Mar. 2021; independent variables of interest are shown in bold.

Panel II: Proportion of trade count; no daily independent variables

	$RTP(OL)_t$ x 100		$RTP(ST)_t$ x 100		$RTP(OC)_t$ x 100		$RTP(MR)_t$ x 100	
	Coef.	t-stat	Coef.	t-stat	Coef.	t-stat	Coef.	t-stat
$Intercept$	−1.2466	−0.70	34.7939***	14.88	16.3045***	7.45	−9.0854***	−9.69
$log(RC_{t-1})$	**0.6426*** **	**7.78**	**0.4166*** **	**8.60**	**1.4949*** **	**17.35**	**−0.0543**	**−1.39**
RTP_{t-1}	74.0138***	57.47	60.5069***	33.83	29.6790***	13.18	71.7417***	55.34
$log(Volume_{t-1})$	1.0441***	7.90	−0.1246	−1.34	0.9686***	5.67	1.0469***	13.31
R_{t-1}	0.6961	0.29	3.9753***	3.67	−7.0611*	−1.89	3.6432**	2.05
Adj. R^2	0.76		0.51		0.35		0.61	
Obs	3556		3556		3554		3556	

$^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$

Results for the same regression with trade count-based RPT measures yields similar results. However, the coefficients are slightly smaller for $RPT(OL)$ and $RPT(ST)$ while the predictive effect of $RPT(OC)$ is larger for trade count-based retail proportion. We suppose that trade count-based measures for stock data are more heavily influenced by a huge number of tiny algorithmic high-frequency trades, adding to the noise and increasing the autocorrelation and thus dwarfing the comment effect. Nevertheless, all three measures are highly significant again. For $RTP(MR)$ we don't get a significant result; we already noted in chapter 4.3.2 that the relationship between the number and volume of retail trades exhibit different characteristics than for our other measures (trades labeled as institutional have a smaller size on average than retail trades for the MR retail classification) and presume that this also affects regression results.

Our second model specification incorporates additional lagged control variables for the previous trading day and time-fixed effects for each half-hour period and weekday.

$$
\begin{aligned}
RTP_t = \alpha + \beta_1 log(1 + RC_{t-1}) + \beta_2 log(1 + dRC_{t-1d}) + \beta_3 RTP_{t-1} + \beta_4 dRTP_{t-1d} \\
+ \beta_5 log(1 + Volume_{t-1}) + \beta_6 log(1 + dVolume_{t-1d}) + \beta_7 R_{t-1} + \beta_8 dR_{t-1d} \\
+ Time + \epsilon \quad (4.2)
\end{aligned}
$$

The additional independent variable of interest, $log(1 + dRC_{t-1d})$ and the previous-day control variables (with preceding d for time $t - 1d$) always relate to the previous trading day.[10] Apart from the variable lag and the different timeframe, the definition of variables stays the same.

[10] E.g. if half-hour t is on a monday, $t - 1d$ relates to the previous Friday.

Table 4.5:
How Reddit Comments affect Retail Trading - Regression Estimates III

This table reports the regression estimates, where the dependent variables are different proxies for the Retail Trading Proportion RTP. RTP measures the proportion of retail trading volume to all trading volume in a 30-minute window. The variants are: i) $RTP(OL)$, the proportion of oddlot-trades; ii) $RTP(ST)$, the proportion of small trades; iii) $RTP(OC)$, the proportion of one-contract option trades and iv) $RTP(MR)$, the proportion of marketable retail orders as defined by Boehmer et al. (2021). All independent variables are measured one period ahead of the dependent variable. The independent variables of interest are: i) RC_{t-1} which denotes the number of all Reddit Comments during the preceding 30-minute window in the WallStreetBets, Stocks and Investing subreddits, where GameStop is explicitly mentioned (either in the post itself or in the parent post) and ii) dRC_{t-1d}, which is defined in the same way as i) but counts all comments made on the previous day instead of the preceding 30-minute window. Additionally, the following control variables are employed, each measured for the preceding 30-minute window and additionally for the previous trading day: i) RTP, a lagged measure of the dependent variable; ii) $Volume$, the trading volume in GameStop shares; iii) R, the absolute return of GameStop shares as a measure for idiosyncratic volatility. Additionally we add Time Fixed Effects (for all 30-minute windows and weekdays) in specifications 2 and 3. We use the Newey and West (1987) methodology with 11 lags to correct standard errors for potential heteroskedasticity and autocorrelation. The sample period is Jan. 2020 - Mar. 2021; independent variables of interest are shown in bold.

Panel III: Proportion of trade volume; with daily independent variables

	$RTP(OL)_t$ x 100		$RTP(ST)_t$ x 100		$RTP(OC)_t$ x 100		$RTP(MR)_t$ x 100	
	Coef.	t-stat	Coef.	t-stat	Coef.	t-stat	Coef.	t-stat
$Intercept$	25.7945***	9.29	6.4148	1.57	10.6238***	5.11	−7.0314***	−3.34
$log(RC_{t-1})$	**0.6258*** **	**4.41**	**0.0754**	**0.51**	**0.4853*** **	**4.66**	**0.0560**	**0.55**
$log(dRC_{t-1d})$	**0.2305** **	**2.11**	**0.2494** **	**2.14**	**0.2013*** **	**2.73**	**0.0436**	**0.56**
RTP_{t-1}	17.7431***	9.33	38.6020***	16.88	11.5929***	4.44	30.4240***	16.10
$dRTP_{t-1d}$	50.0485***	18.54	31.2017***	13.58	42.7518***	8.71	33.6756***	13.43
$log(Volume_{t-1})$	−0.8011***	−4.37	−0.3417	−1.31	−0.5686***	−3.61	1.1377***	6.97
$log(dVolume_{t-1d})$	−0.0436	−0.79	−0.0892	−1.59	−0.0280	−0.73	−0.0826**	−2.01
R_{t-1}	7.3916*	1.94	3.1850	0.59	8.8519***	3.99	−5.1616**	−2.05
dR_{t-1d}	5.8753***	4.40	7.9437***	4.79	0.0561	0.05	0.0579	0.06
Time Fixed Effects	*yes*		*yes*		*yes*		*yes*	
Adj. R^2	0.48		0.42		0.20		0.34	
Obs	3532		3532		3530		3532	

$^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$

Results for volume-based RTP measures can be found in table 4.5. For dependent variables $RTP(OL)$ and $RTP(OC)$ the effect of Reddit comments in the preceding half-hour on retail trading proportion remains highly significant. For both variants and also for $RTP(ST)$ the number of WSB comments on GameStop on the previous trading day additionally significantly affects retail trading activity, while results for $RTP(MR)$ stay below the significance threshold.

Table 4.6:
How Reddit Comments affect Retail Trading - Regression Estimates IV

Table 6 reports the regression estimates, where the dependent variables are different proxies for the Retail Trading Proportion RTP. RTP measures the proportion of the number of retail transactions to all transactions in a 30-minute window. The variants are: i) $RTP(OL)$, the proportion of oddlot-trades; ii) $RTP(ST)$, the proportion of small trades; iii) $RTP(OC)$, the proportion of one-contract option trades and iv) $RTP(MR)$, the proportion of marketable retail orders as defined by Boehmer et al. (2021). All independent variables are measured one period ahead of the dependent variable. The independent variables of interest are: i) RC_{t-1} which denotes the number of all Reddit Comments during the preceding 30-minute window in the WallStreetBets, Stocks and Investing subreddits, where GameStop is explicitly mentioned (either in the post itself or in the parent post) and ii) dRC_{t-1d}, which is defined in the same way as i) but counts all comments made on the previous day instead of the preceding 30-minute window. Additionally, the following control variables are employed, each measured for the preceding 30-minute window and additionally for the previous trading day: i) RTP, a lagged measure of the dependent variable; ii) $Volume$, the trading volume in GameStop shares; iii) R, the absolute return of GameStop shares as a measure for idiosyncratic volatility. Additionally we add Time Fixed Effects (for all 30-minute windows and weekdays) in specifications 2 and 3. We use the Newey and West (1987) methodology with 11 lags to correct standard errors for potential heteroskedasticity and autocorrelation. The sample period is Jan. 2020 - Mar. 2021; independent variables of interest are shown in bold.

Panel IV: Proportion of trade count; with daily independent variables

	$RTP(OL)_t$ x 100		$RTP(ST)_t$ x 100		$RTP(OC)_t$ x 100		$RTP(MR)_t$ x 100	
	Coef.	t-stat	Coef.	t-stat	Coef.	t-stat	Coef.	t-stat
$Intercept$	2.1980	1.00	12.7017***	4.54	15.3597***	5.28	0.2965	0.27
$log(RC_{t-1})$	**0.0427**	**0.43**	**−0.0873**	**−1.50**	**0.4291*** **	**3.44**	**−0.0472**	**−0.88**
$log(dRC_{t-1d})$	**0.0702**	**1.02**	**0.1778*** **	**4.25**	**0.4533*** **	**4.62**	**0.0106**	**0.28**
RTP_{t-1}	56.7956***	35.26	50.0339***	24.39	23.5337***	9.89	59.1990***	35.40
$dRTP_{t-1d}$	33.1538***	20.40	31.3488***	15.38	34.4767***	9.55	29.2051***	15.47
$log(Volume_{t-1})$	0.5929***	4.24	0.1178	1.12	0.5482***	2.91	0.4823***	5.98
$log(dVolume_{t-1d})$	−0.0080	−0.22	−0.0276	−1.31	−0.2826***	−4.33	−0.0498***	−2.58
R_{t-1}	−0.0066	0.00	−2.4071*	−1.73	−1.5583	−0.53	3.3966*	1.82
dR_{t-1d}	3.0822***	4.40	2.3012***	5.32	1.9731**	2.53	−0.2178	−0.34
Time Fixed Effects	*yes*		*yes*		*yes*		*yes*	
Adj. R^2	0.81		0.57		0.39		0.71	
Obs	3532		3532		3530		3532	

$^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$

Looking at the share of retail trades as the number of total transactions with our second model in table 4.6, results get partially murky and we can't confirm the significance of Reddit posts for retail trading activity in all cases. However, our RTP measure with the highest and most robust reaction to Reddit WSB posts throughout all our panels, $RTP(OC)$ based on one-contract option trades, also yields highly significant results for this model.

Looking at the control variables, many of the short-term variables for the preceding half-hour window loose significance in this last specification as well. A larger, cross-sectional long term analysis would be necessary to determine the reasons and confirm robustness of our results.

Overall, we get highly significant results in seven out of eight measure-model combinations for the effect of the number of Reddit WSB comments related to GameStop on the volume-based retail trading proportion in GameStop's shares and options. Surprisingly, our new option-based measure $RPT(OC)$ yields the strongest results and is significant at the 1% level not only for the lagged half-hour window but also for the previous trading day window for every model we ran. This result indicates that the number of Reddit comments has an even bigger effect on retail option trades than on retail stock trades.

However, we conjecture that $RPT(OC)$ might not only be a superior measure for retail trading proportion in the case of Reddit and GameStop, but also for other stocks and situations. Retail option trading has seen tremendous growth over the last year due to low-friction app brokers and at the same time, option order flow is less noisy and might often be easier to attribute due to lower levels of algorithmic trading.

4.4.2 Results of the Granger Causality Test

To confirm our results and as an additional robustness check due to prevalent autocorrelation, we perform a Granger Casuality Test. If the result of our regression analysis is correct, the test should confirm an effect of WSB comments on retail trading proportion but not the other way around. We choose 13 lags for the test as a full trading day consists of 13 half-hour windows.

Table 4.7:
Granger Causality Test

Table 7 shows the results of the Granger causality test. RC denotes the number of all Reddit Comments during a 30-minute window in the WallStreetBets, Stocks and Investing subreddits, where GameStop is explicitly mentioned (either in the post itself or in the parent post). RTP measures the proportion of retail trading volume to all trading volume in a 30-minute window in Panel I and the proportion of the number of retail transactions to all transactions in a 30-minute window in Panel II and is calculated in 4 variants. The variants are: i) $RTP(OL)$, the proportion of oddlot-trades; ii) $RTP(ST)$, the proportion of small trades; iii) $RTP(OC)$, the proportion of one-contract option trades and iv) $RTP(MR)$, the proportion of marketable retail orders as defined by Boehmer et al. (2021). H_0 for the first 2 columns is that the logarithm of RC_{30m} does not Granger cause RTP and the opposite for the last 2 columns. The sample period is Jan. 2020 - Mar. 2021; as one trading day consists of 13 30-minute periods, 13 lags are used for the test.

H_0	$log(RC) \nrightarrow RTP$		$RTP \nrightarrow log(RC)$	
	F-stat	p-value	F-stat	p-value
		Panel I: Trade Volume		
$RTP(OL)$	5.7638***	0.0000	1.3816	0.1597
$RTP(ST)$	3.2713***	0.0001	2.6692***	0.0010
$RTP(OC)$	6.5367***	0.0000	0.7575	0.7062
$RTP(MR)$	2.9979***	0.0002	1.0278	0.4207
		Panel II: Trade Count		
$RTP(OL)$	4.7050***	0.0000	1.7711**	0.0419
$RTP(ST)$	5.9633***	0.0000	2.6111***	0.0013
$RTP(OC)$	7.7191***	0.0000	1.1751	0.2909
$RTP(MR)$	1.6765*	0.0592	1.7732**	0.0415

$^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$

Results can be found in table 4.7. It can be seen that the null hypothesis

$$H_0 : log(RC_{t-1}) \nrightarrow RTP$$

can be discarded with significance at the 1% level for all specifications except for $RTP(MR)$ for trade count proportion, which is only significant at the 10% level. The mirrored relationship is only significant for $RTP(ST)$ for volume-based proportion and weakly significant for all measures except $RTP(OC)$ in the trade-count based panel. Notably, $RTP(OC)$ which showed the strongest and most robust effect accross our regression specifications also achieves the highest F-statistics for the effect of WSB comments on retail trading proportion in both panels in our Granger Causality Test while H_0 can't be discarded for any effect of RTP on WSB comments.

4.5 Conclusion

In this article, we use the unique situation that arose around the GameStop share during 2020 and 2021 with highly elevated trading and investor attention to establish a link between social media activity and the retail trading proportion in shares and options of a company.

Due to the unprecedented social media activity on Reddit's r/WallStreetBets board with more than four million posts directly related to GameStop in our dataset, our results are not only robust over different classifications of retail trades but also significant for high-frequency intraday data. According to our results, a 50% increase in posts on r/WallStreetBets led to an increase in the proportion of retail transactions of up to 0.7% in the following 30 minutes for the shares and up to 0.6% for options. Preliminary results show that the relationship seems to be even stronger in times of very high volatility but further subgroup analysis was out of the scope of this paper.

On January 28th 2021, most brokers partially stopped or limited the opening of new GameStop positions for retail clients[11]. We suppose that RTP would've risen even stronger without this external event, as Reddit posts where at a record high during this time. However,

[11] As can be seen with declining RTP values around that time in figure 4.2.

the magnitude of this effect is difficult to estimate without broker-level individual trade data.

In addition to the strong evidence of social-media induced retail trading, we also were able to compare different measures for classifying retail trades in an extraordinary environment. While all measures display relatively high correlation and seem to be suited to identify partially overlapping shares of the total retail trading volume, conservative measures based on marketable orders only like that from Boehmer et al.'s (2021) seem to be difficult to apply in some situations characterized by extremely high idiosyncratic volatility.

Our newly introduced measure for retail trading based on one-contract option trades on the other hand seems to capture "Reddit-like" retail flow better than traditional stock-based measures. Further cross-sectional analysis and a cross validation of this measure with individual-level broker data would be necessary to confirm whether this holds true outside of this case study as well.

Determinants of Blockholdership *
A new Dataset for Blockholder Analysis

Jan Philipp Harries †

May 2021

Abstract

I introduce a new database for research on institutional and individual blockholders in US-listed firms. Analyzing the contents of all SEC Form 13D and 13G filings downloaded from EDGAR, I find that individual blockholdership of listed companies is in decline over the last 15 years. Holders of large blocks prefer companies with medium market cap, a high equity-to-total-assets ratio and a low dividend-payout ratio. The database is made available free of charge to facilitate further research in this area.

Keywords: Blockholdership, Long-term Ownership, SEC Filings, Dataset

JEL Codes: C81, G23, G32

*A revised excerpt of this article was accepted for publication and is forthcoming in the Journal of Economics as „A Dataset for Blockholders in US-Listed Firms“. This article is accessible at `https://doi.org/10.1515/jbnst-2021-0033` and the accompanying dataset can be found at `https://doi.org/10.7910/DVN/61Z64Q`.

† University of Wuppertal, Gaußstraße 20, 42119 Wuppertal, Germany

5.1 Introduction

Regulatory guidelines in the US command that every investor needs to submit a filing, specifically a Form 13D or Form 13G filing, to the SEC when his stake in a publicly listed company exceeds 5%. These filings are published on the SEC Edgar website and are provided free of charge and publicly accessible. However, while single filings can easily be found and accessed via the EDGAR website, the SEC does not provide a database with aggregated information from these filings and the creation of a database by third parties is hindered by the many different shapes of these filings with a varying layout and wording, despite general template supplied by the SEC.

While the SEC demands many forms to be filed in an unified, machine-readable XML or XBRL format, this is not the case for Form 13D and Form 13G filings. This led to a situation where information from a variety of filings is easily available for researchers (e.g. Form 13F filings, which report holdings of institutional money managers or Form 3, 4 or 5 filings which report insider transactions) while the blockholder information contained in the Form 13D and 13G filings is very difficult to access and thus rarely used in research. Some commercial providers (e.g. Factset) offer proprietary databases claiming to contain information from Form 13D and Form 13G filings, however samples taken from these databases indicate incomplete data. Dlugosz et al. (2006) for example stated that *"despite this important role, there is no standardized data set for these blocks, and the best available data source, Compact Disclosure, has many mistakes and biases"*, referring to a different commercial blockholder database.

With this paper, I introduce and release a new dataset for information contained in Form 13D and Form 13G filings. Manually scanning hundreds of different filing formats, I developed a parser that is sufficiently accurate and robust to parse the relevant information out of hundreds of thousands of blockholder filings. As of June 2021, the resulting database contains 758,666 blockholder filings from November 1993 to May 2021, each with 76 fields containing various information, e.g. the reported ownership percentage or the addresses of the filing entity and the subject company.

Additionally, I run a logistic regression analysis to determine the most important determinants of blockholdership. Albeit limited in scope, the results could help blazing a trail for future blockholdership research by pointing out areas of interest in the way companies with or without significant long-term blockholders exhibit different characteristics. My results show that the likelihood of a company being the subject of a blockholder filing is higher for medium-sized companies with low price-to-revenue ratios and comparatively higher equity-ratios that pay out below-average dividends.

Thus, the contribution of this paper is twofold: First, I describe and release a new publicly-accessible dataset containing all SEC blockholder filings which will enable researchers to use this data to augment existing models and tackle new questions in blockholdership research. Second, I provide a starting point for this research with my own empirical analysis.

The remainder of this paper is structured as follows: Chapter 5.2 gives a short literature review on blockholdership, followed by a brief overview of the regulatory framework for blockholder filings in the US in Chapter 5.3. In chapter 5.4, I explain the most important details of the parsing process before introducing the dataset and providing descriptive statistics and figures. Finally, I present the results of the empirical analysis in chapter 5.5 before concluding in chapter 5.6.[1]

5.2 Relevant Literature

As noted in chapter 5.1, while there are many papers which analyze different aspects of blockholdership, few papers discuss data acquisition and data quality at all. One important paper in this space is Dlugosz et al.'s (2006) *Large blocks of stock: Prevalence, size, and measurement*, which provided methodology and data for following researchers (see also chapter 5.3.2). One very recent paper that also discusses data acquisition and is based directly on Form 13D and 13G filings is *Is Blockholder Diversity Detrimental* by Schwartz-Ziv and Volkova (2020). Using this data to establish a new measure for blockholder diversity, they find that firms with diverse blockholders consistently perform worse than firms with more

[1] Some additional results and figures can be found in appendix 5.A; the code used to parse filings is partially shown in appendix 5.B and a description of all fields in the database can be found in appendix 5.C.

homogeneous blockholders.

Apart from literature directly related to the data, there is a long history of blockholdership and ownership research in financial economics. While an extensive literature review is beyond the scope of this paper, I will briefly touch some of the most interesting and impactful papers in the space.

Shleifer and Vishny (1986) established the potential of large (outside) blockholders to improve firm performance by monitoring managers and preventing agency problems. Mehran (1995) also found benefits of large inside shareholders, a result that recently gained importance again with many founder-led technology companies outperforming others. Bushee (2001) showed that an ownership base dominated by short-term-focused institutional investors could have adverse effects on long-term firm performance. Holderness (2003) provided a survey of previous literature on blockholdership. Amongst other results he found no evidence that ownership concentration has an impact on firm value. Contrary to our more recent results in chapter 5.5, he found no significant relationship between blockholdership and leverage in the previous literature.

Andres (2008) showed that firms with family blockholders (and family board representation) are more profitable than firms that are widely-held or have other types of blockholders using a sample of German exchange-listed companies. Cronqvist and Fahlenbrach (2009) established that heterogeneity across large shareholders causes several policy effects and find significant effects on investment, financial and executive compensation policies as well as on firm performance measures. Edmans (2009) introduced a model to prove that even small blockholders lacking control rights can significantly enhance firm value, for example by encouraging managers to take a long-term perspective. Holderness (2009) uses a random sample of US firms and hand-collected blockholder data to show that blockholders in aggregate owned an average of 39% of surveyed companies at that time.

Clifford and Lindsey (2016) offer a new perspective on blockholder heterogeneity, finding that blockholders that actively monitor firms cause greater improvements in operating-performance in these firms than blockholders that are monitoring only passively. Edmans and Holderness (2017) provide an extensive review of theory and empirical findings in the blockholder literature.

More recently, Backus et al. (2019) show that the strong rise in common ownership between 1980 and 2017 is driven primarily by the rise of indexing and diversivcation, Hadlock and Schwartz-Ziv (2018) observe that non-financial blockholder prefer smaller, riskier, younger, and less-liquid firms (which is partially confirmed by my results in chapter 5.5) and Aminadav and Papaioannou (2020) focus on corporate control of blockholders and extend the predominantly US-centric perspective in blockholdership research with data on firms around the world.

5.3 Regulatory Framework and Alternative Data Sources

To explain the relevance of Form 13D and Form 13G SEC filings and enable researchers to better understand possibilities and limits of the released blockholder dataset, I briefly explain the regulatory framework for these filings and point out alternative data sources.

5.3.1 Regulatory Framework for Blockholder Filings in the US

The Securities Exchange Act of 1934 contains the filing rules and legal definitions under which blockholders have to submit forms to the SEC. Relevant for blockholders is first and foremost section 13(d).[2]

Section 13(d)-1(a) commands that "any person who, after acquiring directly or indirectly the beneficial ownership of any equity security of a class which is specified in paragraph (i) of this section, is directly or indirectly the beneficial owner of more than five percent of the class shall, within 10 days after the acquisition, file with the Commission, a statement containing the information required by Schedule 13D". As noted by Dlugosz et al. (2006), "this rule has been interpreted to include shares that may be obtained through the exercising of options, warrants, or rights in the next 60 days a part of the beneficial ownership calculation".[3] These

[2] See also *43 FR 18495* or *17 CFR 240.13d-1* ff. in the US Code of Federal Regulations.

[3] For more information and interpretation of these rules, see also "Exchange Act Sections 13(d) and

filings are known as Form 13D filing.

Some types of investors (broker-dealers, banks, insurance companies, investment companies, investment advisors, employee benefit plans, parent companies, savings associations and churches) are allowed to file an abbreviated filing, the Form 13G, which waives some required fields (e.g. the source of funds). This is only possible if the investor has "has acquired such securities in the ordinary course of his business and not with the purpose nor with the effect of changing or influencing the control of the issuer, nor in connection with or as a participant in any transaction having such purpose or effect". The exact conditions are laid out in section 13(d)-1(b) and these forms were also parsed for the database.

In the case of a change in a previously reported ownership share in a company, the (former) blockholder has to file an amendment to the original filing. These filings are called Form 13D/A or Form 13G/A respectively.

All forms are to be published on the SEC's EDGAR platform[4]. Besides Form 13D and G filings, other filings of interest for company ownership are for example the Forms 3, 4 and 5, which are used to disclose insiders transactions and Form DEF 14A proxy statements. Further information regarding other filings and forms can be found e.g. at Meredith (2007).

5.3.2 Alternative Data Sources

Closest to my dataset is the recently-released blockholder dataset by Schwartz-Ziv and Volkova (2020), which was also compiled direcetly from Form 13D and 13G filings. While their database is fundamentally based on the same filings and also a great resource for researchers, it comes with several limitations: 1) It contains only data until 2015 while I provide data until May 2021. 2) It is aggregated on a year-filer-subject level, which means that additional filings are discarded. 1) and 2) lead to a reduced size of 389,818 filings in their database compared to my 758,666 available filings. 3) It contains only four columns (CIKs of blockholders and subject companies, the interpolated year of the filing and the

13(g) and Regulation 13D-G Beneficial Ownership Reporting" at https://www.sec.gov/divisions/corpfin/guidance/reg13d-interp.htm.

[4] The platform can be accessed at https://www.sec.gov/edgar/searchedgar/companysearch.html.

ownership percentage) compared to 76 columns in my database. This not only limits the scope of addressable research (e.g. because field like investor type or source of funds are missing), but also prevents a better matching of companies (In my database, CUSIP and names are available for all filers and subject companies instead of only CIKs.). 4) It misses some filings due to unknown reasons[5]. However, I was able to confirm the correctness of parsed block percentages for a random sample of entries in their database and due to the difficult parsing of heterogeneous filings, it will be beneficial for researchers to have access to multiple data sources in any case.

The Factset ownership database is also commonly used when analyzing blockholder behaviour (e.g. Hadlock and Schwartz-Ziv, 2018). While Factset claims to also use data from Form 13D and Form 13G filings, I found many instances of missing filings in publicly available sample data.[6] Other researchers use S&P's Compact Disclosure database, which has known weaknesses and only extends through 2006 (see e.g. Clifford, 2008), use a subset of firms with manually collected blockholders, often based on the methodology by Dlugosz et al. (2006) (see e.g. Cronqvist and Fahlenbrach, 2009; Edmans and Holderness, 2017), or limit their analysis to blocks contained in Form 13F filings, which are generally easier to access (see e.g. Anton et al., 2016; Gloßner, 2019). While I don't have access to all alternative data sources to compare them in detail, I assume that currently no other publicly-available, non-commercial and complete database of parsed Form 13D and Form 13G blockholder filings is available for researchers.

[5] E.g. this one `https://www.sec.gov/Archives/edgar/data/0000899749/000091420813000064/0000914208-13-000064-index.htm` or this one `https://www.sec.gov/Archives/edgar/data/105982/0000105982-94-000049.txt`.

[6] E.g. 0 of 6 Form 13D or 13G filings for the company LiveRamp (formerly ACXIOM) in 2018 and 0 of 3 Form 13D or 13G filings for AMETEK in 2017/18. In contrast, data from easier-to-parse Form 13F, Form 3, 4, 5 and Proxy statements seems to be very reliable at a cursory glance.

5.4 Explorative Analysis of Form 13-D and Form 13-G Filings

In this chapter, I will first describe the process of mining and parsing filing data in text-form including some of the challenges which have an impact on data quality before sharing some descriptive statistics on the resulting dataset and presenting the finished database.

5.4.1 Parsing of Blockholdership Filings

Before I create the database of blockholdership filings, the filings need to be downloaded and parsed. I downloaded 758,666 raw text-form 13-D, 13-D/A, 13-G and 13-G/A filings from EDGAR spawning from November 1993 to May 2021 using the EDGAR master-files, containing links to all released filings. A random sample of filings, which was manually checked, did not yield any missing filings.

Standard filings start with a header that is structured as in the following example[7]:

```
<SEC-DOCUMENT>0000353296-99-000011.txt : 19990126
<SEC-HEADER>0000353296-99-000011.hdr.sgml : 19990126
ACCESSION NUMBER:                 0000353296-99-000011
CONFORMED SUBMISSION TYPE:         SC 13G
PUBLIC DOCUMENT COUNT:                 1
FILED AS OF DATE:                 19990125

SUBJECT COMPANY:

        COMPANY DATA:
                COMPANY CONFORMED NAME:                         SAVILLE SYSTEMS PLC
                CENTRAL INDEX KEY:                         0001001635
                STANDARD INDUSTRIAL CLASSIFICATION:         PROGRAMMING SERVICES [7371]
                IRS NUMBER:                                 000000000
...
```

Document link, accession number, submission type, document count, date and data on the subject company and the filing entity can be extracted from this header. However, the most

[7] See `https://www.sec.gov/Archives/edgar/data/353296/0000353296-99-000011.txt` for this example filing.

important part follows later in the document:

```
11. Percent of Class Represented by amount in #9
      5.1
```

This part contains the important information about the size of the ownership stake. However, as laid out before, the Form 13D and 13G filings are unfortunately not formatted in a machine-readable format and even if the SEC supplies template for the form layout (e.g. the ownership stake is always reported as item 11 or 13), there are multiple different, difficult-to-parse variations in the data. Examples for problematic format are padding symbols like in the following filing which could also be interpreted as 0.22%[8].

```
(13) PERCENT OF CLASS REPRESENTED BY AMOUNT IN ROW (11)...........22%
```

or filings that hide the percentage between HTML-tags like[9]

```
....TEXT-INDENT: 0pt; TEXT-ALIGN: left"><font size="2">100%</font></p>
```

These are just a few examples that highlight how correctly parsing the filings can be a challenging task. In the end, I decided to implement an heuristical, multi-tiered approach, which yielded by far the best results compared to simpler approaches with manual parser-selection.

Initially, the parser splits header and body (using a fallback method if correct tags aren't found) of the filing. In the next step, it parses all information contained in the header (which is less challenging than information contained in the body due to a more homogeneous format). Afterwards, the body is separated into multiple documents (as one filing can contain multiple documents) and documents belonging to the correct filing type are parsed for all important information (mainly ownership percentage, group membership, investor type, source of funds and several legal properties).

Taking ownership percentage as an example, the parser then splits the content of the document into an array of single lines of text. The advantage of this approach is that features like proximity are easier to calculate when searching for the correct number that denotes the

[8] The original filing can be found at https://www.sec.gov/Archives/edgar/data/1000366/0000950129-99-000206.txt.

[9] The original filing can be found at https://www.sec.gov/Archives/edgar/data/1000275/000107330709000008/0001073307-09-000008.txt.

ownership percentage. In the next step, the parser searches for all lines containing the word "percent" (as this word appears in the template provided by the SEC next to this item). The lines following these occurrences are then marked as possible locations of the percentage the parser is searching for. Afterwards, these possible locations are passed on to a method that extracts all percentage-candidates in two steps[10]:

1. Search for any possible percentage values (e.g. 0.000-100.0) preceding a percentage sign.
2. Search for any other numbers that could be percentages in the candidate lines if nothing was found in step 1.

The numbers are then extracted using the following regular expressions, which are the results of manual tests with hundreds of filings:

```
#with percentage sign
(^|\s|(\.\.)|:|=)\d{1,3}([\.\,]\d{1,3})?\s*(?=%)
#without percentage sign
(^|\s|(\.\.)|:|=)\d{1,3}[\.\,]\d{1,3}($|\s)
```

In the next step, the parser checks for all candidates found whether they are float numbers (containing a dot) or integers (just natural numbers). Due to many factors (e.g. HTML-attributes, enumerations), the searched-for ownership percentage is more likely to be a float number. However, there are also filings containing only integer numbers and thus discarding all integers from the parser would lead to incorrect results. If any floats are found, all integers are discarded and the parser picks the largest found candidate[11]. If only integers are among the candidates, ambiguous numbers (e.g. 5, 9, 10, 11 which appear in almost every filing text) are filtered as long as other candidate percentages are still available. Combining these techniques, I achieve a high accuracy, judging by comparisons with other data sources and manual checks.

[10] Please see appendix 5.6 for the source code.

[11] In more than 99% of cases, there is only one candidate left. However, in rare cases containing multiple entities or share classes, we usually want to pick the largest available percentage number, as smaller percentages may refer e.g. to subsets of the reported percentage split by sub-entity or share class.

Figure 5.1 shows the share of all parsed filings where no valid value for the respective category could be extracted.

In the following chapter 5.4.2, I will show that the parser yields good and plausible results in the overwhelming number of cases and provide descriptive statistics for parsed filings.

5.4.2 Dataset and Descriptive Analysis

The final database contains 758,666 unique blockholdership filings, thereof 205,063 Form 13G filings, 374,531 Form 13G/A filings, 57,122 Form 13D filings and 121,950 Form 13D/A filings. The development of the number of the different filing types over time can be seen in figure 5.2. Since 2008, the number of filings, especially Form 13 D filings, has declined slowly but steadily.

There are 28,246 unique CIK's among the filing subjects and 59,628 unique CIK's among the filers themselves in the database. The list of subjects with most filings is dominated by several ETF's. *iShares ETF's* were most reported on with 2.206 filings, followed by *WM Advisors* with 667 filings and *IndexIQ ETF Trust* with 649 filings (see table 5.1 for full results).

Table 5.1: Top 5 Filing Subjects

This Table lists the top five filing subjects (Subject is the company for which the blockholder declares his stake). Period: November 1993 to May 2021.

Filing Subject	# Filings	Median %
iSHARES TRUST	2266	11.30
WM ADVISORS INC	667	12.10
IndexIQ ETF Trust	649	57.38
ESTEE LAUDER COMPANIES INC	396	5.50
STEELCASE INC	390	7.20

Among Filers, data is more concentrated. Leading the pack is *BlackRock Inc.* with 34,253 filings (which amounts to almost 5% of all filings) followed by *VANGUARD* with 21,059 filings and two different entities belonging to *Fidelity* which, if taken together, even top

Figure 5.1: Share of missing Values for scraped Items

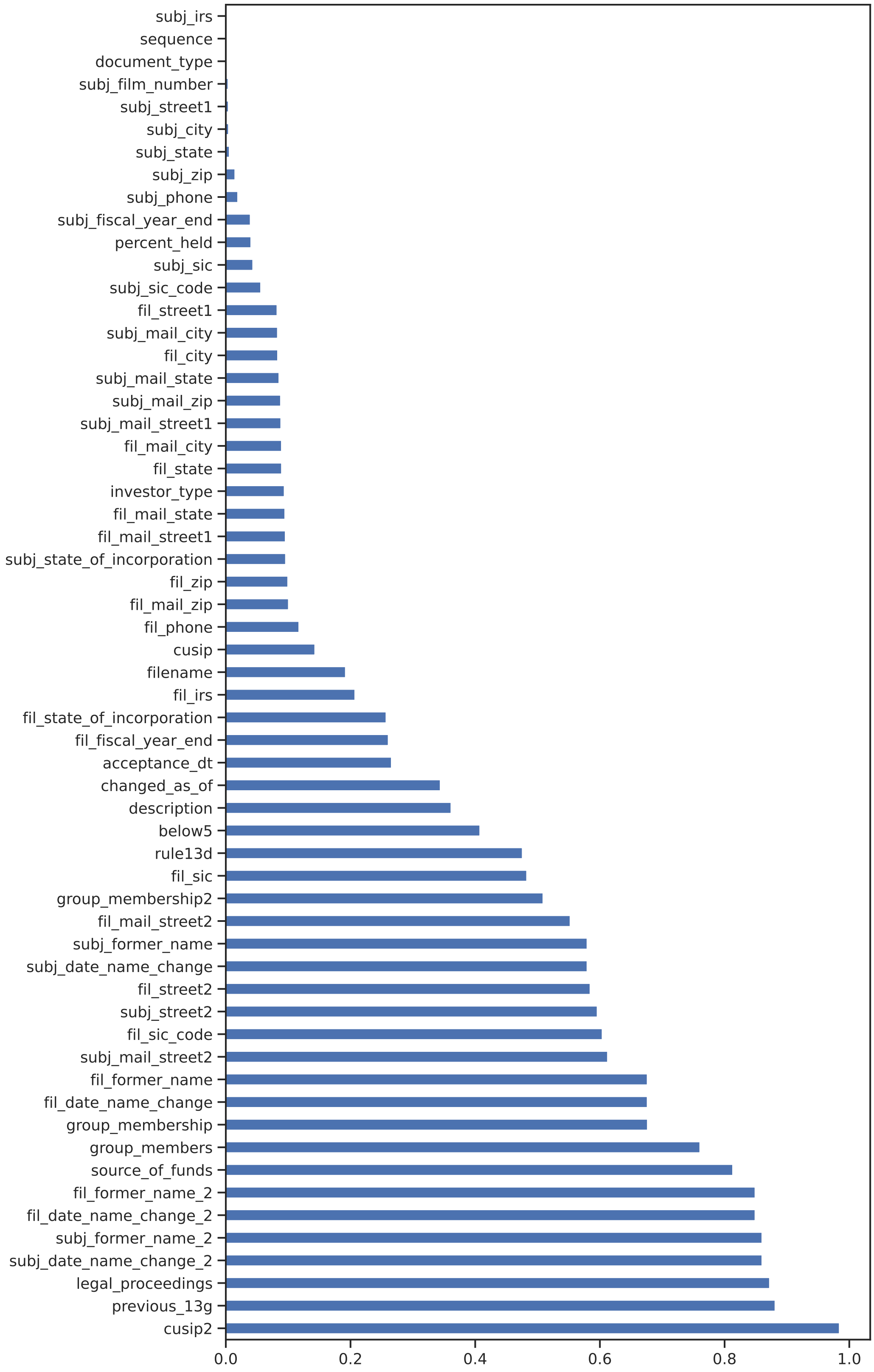
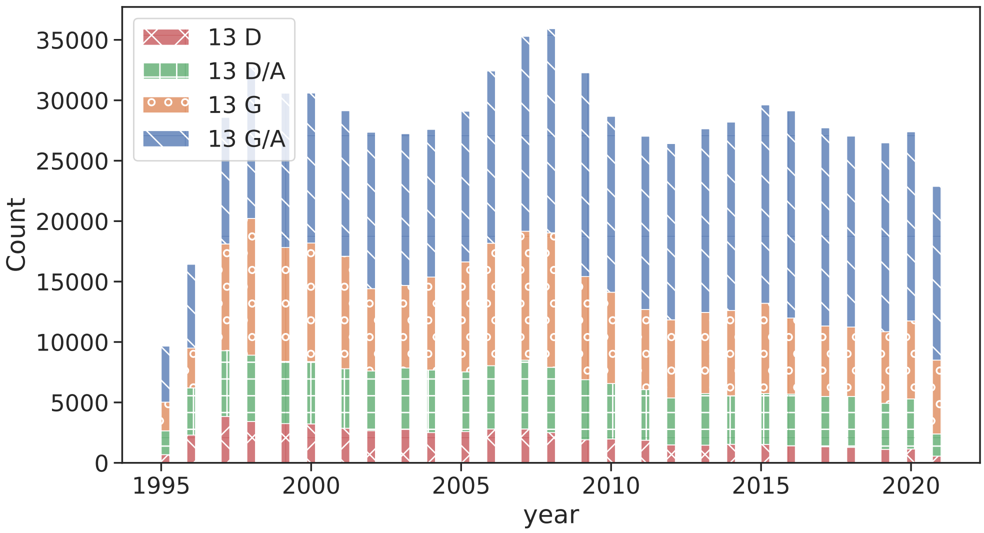


Figure 5.2: Filing Count per Year and Filing Type



BlackRock with a combined total of 37,769 filings. All big filers report a median blocksize around seven or eight percent.

Geographically, most subject companies are located in New York, followed by Chicago, San Diego and San Francisco (see table 5.3).

While New York also takes the top spot for the location of the filing entity (with a total of 161,266 filings over all Zip codes), some cities like Boston and Baltimore have significantly more local filers than subjects, probably indicating a bigger financial industry based in these locations. The top five filer ZIP codes can be found in table 5.4.

Table 5.2: Top 5 Filers

This Table lists the top five filers (The entity which declares a stake). Period: November 1993 to May 2021.

Filing Subject	# Filings	Median %
BlackRock Inc.	34 253	7.30
VANGUARD GROUP INC	21 059	7.45
FMR CORP	20 789	8.09
FMR LLC	17 080	8.17
PRICE T ROWE ASSOCIATES INC	13 139	7.50

Table 5.3: Top 5 Locations of Subject Companies

This Table lists the top five ZIP codes of filing subjects (Subject is the company for which the blockholder declares his stake). Period: November 1993 to May 2021.

Zip Code	# Filings	City, State
10022	6627	New York, NY
60606	5210	Chicago, IL
92121	5012	San Diego, CA
94105	4977	San Francisco, CA
10019	4407	New York, NY

Table 5.4: Top 5 Locations of Filers

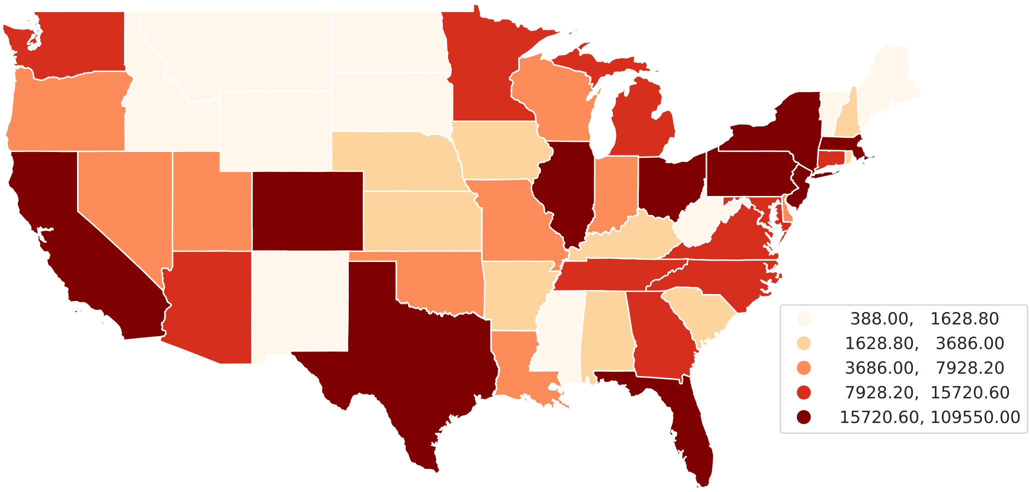
This Table lists the top five locations of filers (The entity which declares a stake). Period: November 1993 to May 2021.

Zip Code	# Filings	City, State
02109	42 892	Boston, MA
10055	31 465	New York, NY
10022	30 048	New York, NY
10019	17 951	New York, NY
02210	15 792	Boston, MA

If looking visually at the distribution of filing entities and subject companies, this picture is confirmed. The largest US states all fall in the highest quintile of filings as can be seen in figure 5.3. California is the state where most subject companies are based with a total of 109,550 filings. For filers, New York takes unsurprisingly the top spot followed by California (85,971 filings) and Massachusetts (81,645 filings). As can be seen from figure 5.4, states that don't contain financial hubs like e.g. Florida or Colorado will be found more often as the location of the subject company than location of filing entity.

The histogram of reported ownership (measured in % of outstanding shares) in figure 5.5 shows that only very few filings report ownership stakes larger than 20%. The overwhelming majority of filings reports a share of 10% or less of a company.

Figure 5.3: Geographical Distribution of Subject Companies

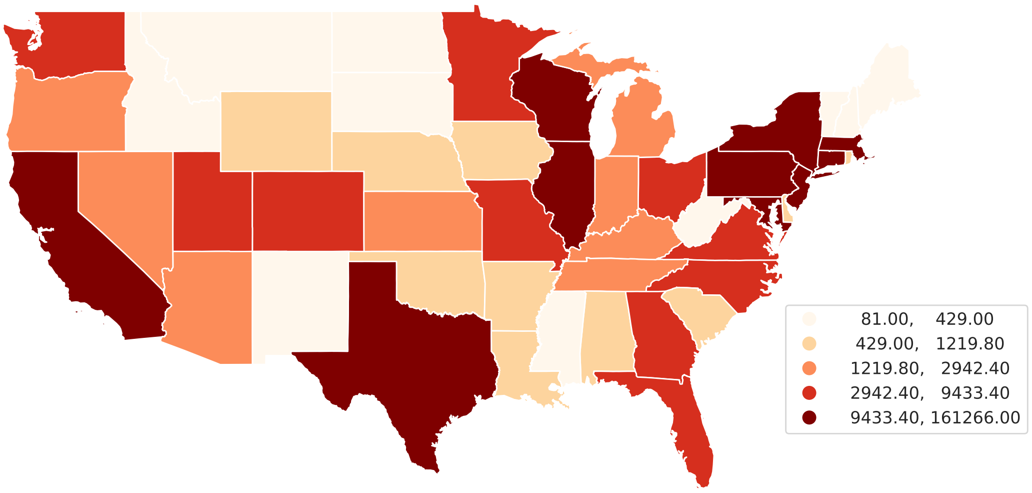


As can be seen in figure 5.6, reported ownership percentages have stayed remarkably similar over time. However, the 75% quantile of reported ownership has come down since the early 2000's from about 13% to 10% in 2021.

When filing a Form 13D or 13G with the SEC, investors have to self-report their investor category from a choice of presets. Investors from different categories exhibit different characteristics in their number of filings and the average reported blocksize, as can be seen in table 5.5. Most filings are filed by either Investment Advisors or Individuals with 211,711 and 131,474 total filings respectively. Interestingly, the size of the average reported ownership percentages differs significantly by investor type. While Investment Advisors have among the lowest average reported ownership percentage with 7.85%, Individuals report an average ownership that is two times as high with 15.65% which is only topped by Corporations with an average of 20.56%.

Figure 5.7 shows an interesting trend in the data, which might be explained by the recent trend to passive investments, the increasing market capitalization of listed companies and more concentrated ownership in general. While the number of filings from individuals and investment advisors was relatively similar until 2005, since then the development is very stable for investment advisors with about 8,000 filings so far in 2021, while the number of filings from individuals declined sharply from about 7,000 in 1998 to only 2,000 in 2021.

Figure 5.4: Geographical Distribution of Filers



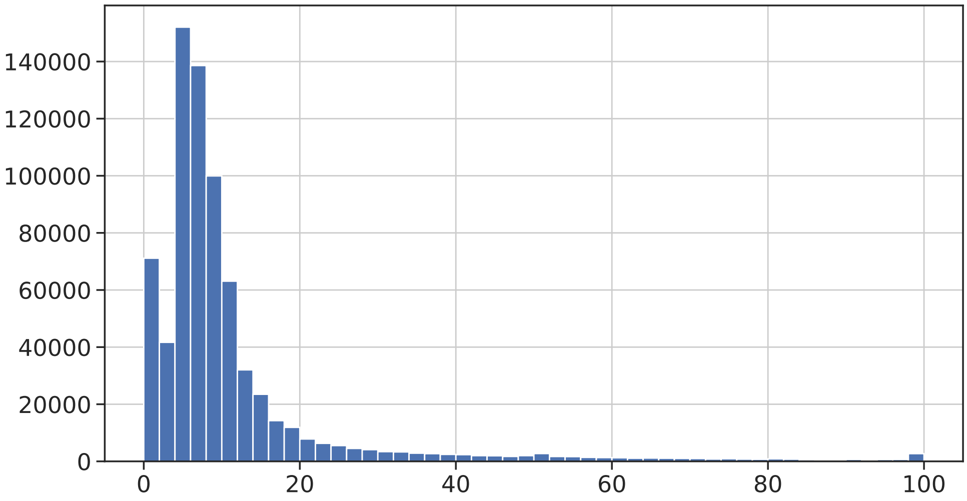
For Form 13D and 13D/A filings, the filer has also to declare the source of funds, e.g. how he financed the purchase of the reported blockholding. Descriptive statistics for this category are given in table 5.6. The average reported ownership size is biggest for *Bank* as funding source with an average of 36.01%. It's not immediately clear why the average size for reported bank-funded blocks is larger than for the other categories. One possible explanation could be that M&A transactions are often the reason for a large reported ownership share and these transactions are often (at least partially) financed by banks.

Among the other fund sources, blocks financed by *Affiliates*, *Personal Funds* and *Working Capital* exhibit a below-average size. Categories with most filings are *Working Capital* and *Other*.

5.5 Empirical Analysis of the Determinants of Blockholdership

In this chapter, I will use the dataset to empirically analyze determinants of blockholdership. This analysis results in relevant new insights and also confirms some of the prior literature on the topic. However, in some cases the empirical analysis is still comparatively shallow in relation to the research opportunities enabled by this dataset. Thus, the main contribution of

Figure 5.5: Histogram of reported Ownership Percentages



this analysis is paving the way for following research, tackling the more challenging questions and implications of blockholdership and company ownership.

5.5.1 Data and Variables

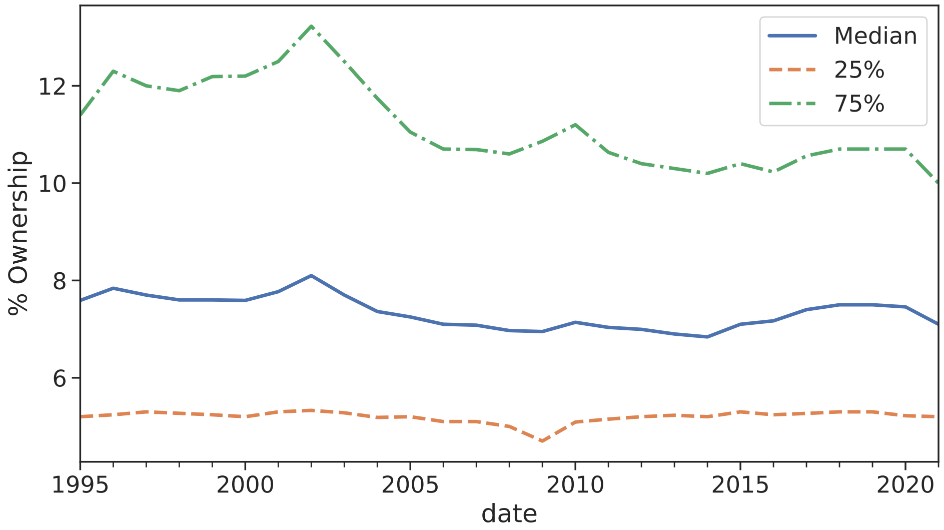
To get a better understanding of the determinants of blockholdership, e.g. in which ways companies with (large) blockholders differ from companies with a comparatively higher public float, I merge the blockholdership database (for more information on that data see the previous chapter 5.4.2) with annual fundamental and financial data sourced from Compustat. The Compustat data consists of 322.596 firm years (from 1993 to 2020) on 24.410 US-listed companies. To limit the effect of outliers and reporting errors, raw Compustat data is winsorized at the 1% and 99% level.

By matching via unique CIK and CUSIP identifiers and aggregating all reports, I find 143.333 firm years with matching blockholder reports. Companies with matched blockholder reports have a median of 3 reported blocks while the median largest reported block in a given firm-year is 10.70 % and the cumulative percentage for all reported blocks is 25.98 %.

For my empirical analysis, I construct the following independent variables:

- $log(MCap)$: The logarithm of the company's market capitalization, calculated as prod-

Figure 5.6: Development of reported Ownership Percentages



uct of Compustat fields *CSHO* and *PRCC_C*.

- P/B: The price-to-book ratio, calculated as the ratio of market capitalization to book equity[12].
- P/E: The price-to-earnings ratio, calculated as the ratio of market capitalization to net income (Compustat code *NI*). Negative P/Es are excluded.
- P/R: The price-to-revenue ratio, calculated as the ratio of market capitalization to total revenue (Compustat code *REVT*).
- ER: The equity ratio, calculated as the ratio of the book value of common equity (Compustat code *CEQ*) to total asset (Compustat code *AT*). Negative ERs are excluded.
- DR: The dividend-payout ratio, calculated as the ratio of dividend distributions (Compustat code *DV*) to market capitalization.

[12] For the calculation of book equity, I use the procedure laid out by Fama and French (1995), using the book value of stockholder's equity plus balance-sheet deferred taxes and investment tax credit minus the book value of preferred stock.

Table 5.5: Summary Statistics: Investor Type

This table reports summary statistics for the ownership size (as % of floating shares) reported in Form 13D and 13G (/A) filings by self-reported investor category. Period: November 1993 to May 2021.

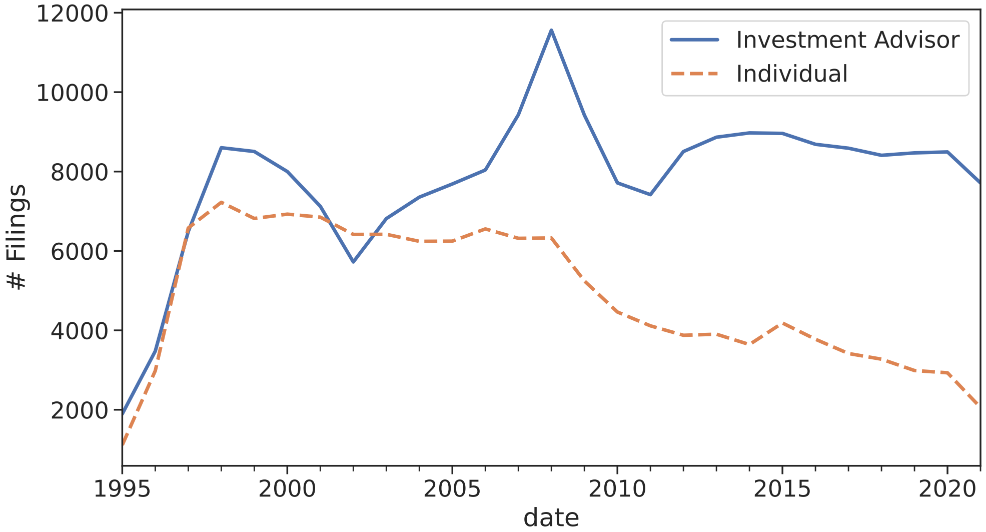
Investor Type	N	Mean	Std	Min	Q1	Median	Q3	Max
Bank	11 545	9.54	13.50	0.00	4.31	6.18	9.31	100.00
Broker	11 061	9.80	9.72	0.00	5.30	7.48	11.30	100.00
Church	13	15.83	13.66	0.00	5.80	11.12	19.99	49.23
Corporation	53 254	20.56	23.59	0.00	5.52	9.99	26.60	100.00
Employee Benefit	7969	9.89	8.28	0.00	5.92	8.40	11.17	100.00
Holding Company	95 178	9.56	12.17	0.00	5.17	6.76	9.99	100.00
Individual	131 474	15.65	17.95	0.00	5.70	9.00	17.60	100.00
Insurance	7145	10.61	13.98	0.00	4.96	6.88	10.80	100.00
Investment Advisor	211 711	7.85	7.31	0.00	5.18	6.70	9.49	100.00
Investment Company	17 359	8.98	8.29	0.00	5.50	7.42	10.46	100.00
Non-US Institution	3807	6.84	9.62	0.00	3.30	4.74	7.20	100.00
Other	56 062	13.37	17.34	0.00	5.00	7.40	13.60	100.00
Partnership	66 248	11.03	13.94	0.00	4.40	6.90	11.30	100.00
Savings Association	158	7.56	4.79	0.00	5.03	6.27	8.79	43.60

- *GW/TA*: The goodwill-to-assets ratio, calculated as the ratio of the value of Goodwill on the balance sheet (Compustat code *GDWL*) to total assets (Compustat code *AT*).

The choice of independent variables reflects the goal to learn more about the characteristics of companies with long-term or large-stake blockholders and in which ways these companies differ from companies without these blockholders. Previous results in the literature and theoretical considerations (see chapter 5.2) imply that blockholders are more interested in long-term profitable companies, so that it would be fair to assume a lower valuation relative to earnings and book value, lower dividend-payout (due to more focus on long-term investments) and higher equity-ratio for companies with blockholder reports.

Dependent variables used in our analysis are the following (each set to zero or one to indicate a firm-year with the specified property):

Figure 5.7: Development of Filings filed by Investments Adivsors and Individuals



- $BH(any)$ indicates that the company has at least one blockholder who reported any ownership stake in the corresponding firm-year.

- $BH(o10)$ indicates that at least one blockholder has declared a stake of equal or above 10% in the company in the corresponding firm-year.

- $BH(t33)$ indicates that the ownership stake reported by all blockholders together for the firm-year is higher than 33%.

These dependent variables are used in two variants: All blockholders are counted in Panels I and III, while in Panels II and IV only non-financial blockholders are counted (excluding blockholders who self-identify as Bank, Broker, Insurance, Investment Advisor, Investment Company or Savings Association). As the literature on financial blockholders is comparatively extensive due to the good availability of Form 13F holdings data, the focus on non-financial holders in Panel II and IV will help to shine a light on this under-represented group of blockholders.

After dropping firm-years with missing data, 101,229 firm-years are available to use in the following regression analysis. Summary statistics for these dependent and independent variables can be found in table 5.7.

Table 5.6: Summary Statistics: Source of Funds

This table reports summary statistics for the ownership size (as % of floating shares) reported in filings by self-reported source of funds category. Note that the number of observations is limited as source of funds has only to be reported for Form 13D and 13D/A filings. Period: November 1993 to May 2021.

Source of Funds	N	Mean	Std	Min	Q1	Median	Q3	Max
Affiliate	17 922	19.01	20.72	0.00	6.11	10.10	23.20	100.00
Bank	1572	36.01	28.40	0.00	12.14	28.50	52.23	100.00
Other	43 836	22.92	23.32	0.00	6.64	13.68	31.59	100.00
Personal Funds	26 430	20.65	21.93	0.00	6.40	11.50	26.60	100.00
Subject Company	2090	22.06	20.99	0.00	7.20	13.95	29.76	100.00
Working Capital	47 565	19.62	22.04	0.00	5.72	10.30	24.40	100.00

Additionally, I add Time and Industry Fixed Effects[13] to the regression analysis performed in the following chapter 5.5.2.

5.5.2 Empirical Results

I model blockholdership as dependent from various fundamental ratios and the size of a company. The following logistic regression is used in my analysis:

$$\begin{aligned} P(BH=1) = S(\alpha + \beta_1 log(MCap) + \beta_2 P/B + \beta_3 P/E + \beta_4 P/R \\ + \beta_5 ER) + \beta_6 DR + \beta_7 GW/TA + Time + Industry) \end{aligned} \tag{5.1}$$

with S being the sigmoid function. Z-Values are given along with the coefficients and p-values are estimated with a Wald test.

In table 5.8, results for Panel I (all companies and all blockholders) are shown. With the exception of P/B and P/E, all independent variables are highly significant for all variants of blockholdership BH. The positive coefficient for $log(MCap)$ (which indicates that the likelihood of any reported blockholder is higher for larger companies, which is somewhat counter-intuitive) supposedly comes from a large number of very small listed companies

[13] For industry classification I use the first digit of the SIC-classification.

Table 5.7: Summary Statistics: Indepedent and Dependent Variables

This table reports summary statistics for independent and dependent variables used in the logistic regression analysis. Period: November 1993 to May 2021.

Variable	N	Mean	Std	Min	Q1	Median	Q3	Max
$log(MCap)$	101 229	19.98	2.37	5.92	18.30	19.99	21.61	28.23
P/B	101 229	2.95	4.66	0.17	1.14	1.80	3.07	55.41
P/E	101 229	35.17	69.87	0.88	11.61	17.77	29.27	551.46
P/R	101 229	3.04	15.64	0.02	0.66	1.43	2.92	543.18
ER	101 229	0.46	0.25	0.03	0.27	0.46	0.65	0.99
DR	101 229	0.02	0.03	0.00	0.00	0.00	0.02	0.20
GW/TA	101 229	0.08	0.13	0.00	0.00	0.01	0.12	0.56
Panel I/III: All Blockholders								
$BH(any)$	101 229	0.67	0.47	0.00	0.00	1.00	1.00	1.00
$BH(o10)$	101 229	0.37	0.48	0.00	0.00	0.00	1.00	1.00
$BH(t33)$	101 229	0.26	0.44	0.00	0.00	0.00	1.00	1.00
Panel II/IV: Only non-financial Blockholders								
$BH(any)_{nonfin}$	101 229	0.55	0.50	0.00	0.00	1.00	1.00	1.00
$BH(o10)_{nonfin}$	101 229	0.23	0.42	0.00	0.00	0.00	0.00	1.00
$BH(t33)_{nonfin}$	101 229	0.12	0.32	0.00	0.00	0.00	0.00	1.00

in the dataset, where filings are often missing or can't be merged with fundamental data. As can be seen from Panel III (see table 5.10 in the Appendix), if I restrict companies to companies with a market capitalization above USD 1bn, the coefficient becomes negative for all dependent variables, as originally expected. Overall, medium-sized companies are most likely to have filing blockholders. For $BH(o10)$ and $BH(t33)$, the coefficient is even more negative (while still highly significant with z-values around -45). This shows that it is more likely for small companies (given a minimum market capitalization of USD 1bn) to have blockholders that report blocks above 10% and/or a cumulative reported blockholdership of above 33%.

Results for P/B and P/E in Panel I don't paint a clear picture, however P/R shows a significant negative coefficient for all BH variants. This indicates that companies with a low price-to-revenue ratio are more likely to have reporting blockholders, confirming the hypothesis that blockholders often prefer moderately valued companies. For the equity ratio ER, I find significant positive ratios for all BH variants. The higher the equity ratio of a company, the higher the likelihood of reporting blockholders. The effect is most pronounced for companies with at least one blockholder reporting a stake above 10%, as can be seen from the very high z-value of 19.27. This results could indicate a stronger long-term orientation of blockholders and show their hesitancy to invest in over-levered companies with a large amount of debt on the balance sheet.

Coefficients for the dividend-payout ratio DR are significantly negative for all specifications, meaning that companies with a low dividend-payout ratio have a higher likelihood of reporting blockholders. This makes sense if one considers holders of large blocks to be more interested in the long-term health of the company than in short-term cash distributions. Finally, coefficients for GW/TA are significantly positive for all BH measures, meaning that companies with a high Goodwill-to-assets ratio are more likely to have reporting blockholders. This comes somewhat surprisingly, as one would have expected that holders of larger blocks are overwhelmingly critical of possibly-inflated balance sheets due to a high Goodwill position. However, compared to other measures the coefficients are relatively small and as we will see in Panel II, results for non-financial blockholders are somewhat different.

Table 5.8:
Regression Analysis: Determinants of Blockholdership I

This table reports the logistic regression estimates, where the dependent variables are different proxies for Blockholdership BH. BH is 1 if the respective conditions are met by the company: i) $BH(any)$ is one if the company has at least one blockholder who reported any ownership percentage in the firm-year; ii) $BH(o10)$ is one if at least one blockholder has declared a stake of equal or above 10% and iii) $BH(t33)$ is one if the ownership stake reported by all blockholders together for the firm-year is above 33%. The independent variables are: i) $log(MCap)$, the logarithm of the company's market capitalization; ii) P/B, the price-to-book ratio; iii) P/E, the price-to-earnings ratio; iv) P/R, the price-to-revenues ratio; v) ER, the equity-ratio; vi) DR, the dividend-payout-ratio and vii) GW/TA, the ratio of goodwill to total assets. Additionally, I add Time Fixed Effects (for all years in the sample) and Industry Fixed Effects (for SIC-1). P-values are calculated with a Wald test. Period: November 1993 to May 2021.

Panel I: All companies, All Blockholders

	$BH(any)$		$BH(o10)$		$BH(t33)$	
	Coef.	z-val	Coef.	z-val	Coef.	z-val
$Intercept$	−11.6888***	−11.65	−9.6788***	−9.65	−9.1715***	−9.14
$log(MCap)$	0.1566***	43.81	0.0548***	16.93	0.0265***	7.48
P/B	−0.0121***	−7.34	0.0032**	2.08	0.0005	0.30
P/E	−0.0002	−1.64	0.0006***	5.66	0.0008***	7.64
P/R	−0.0034***	−6.92	−0.0026***	−4.74	−0.0036***	−4.80
ER	0.3409***	9.66	0.6308***	19.27	0.4914***	13.59
DR	−3.2761***	−13.85	−2.5818***	−10.62	−3.4134***	−12.23
GW/TA	1.4103***	19.13	0.6898***	12.06	0.6780***	11.22
Time & Industry FE	yes		yes		yes	
Pseudo R^2	0.17		0.08		0.08	
Obs	101229		101229		101229	
$\sum(BH = 1)$	68196		37598		26776	

$^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$

Results for Panel II are given in table 5.9. For this panel, analyzed blockholdership filings are restricted to non-financial blockholders. Similar to Panel I, most variables display a significant effect on the likelihood of blockholdership. While P/E exhibits a small but significant positive coefficient for all BH variants, results for P/B are mixed again, confirming the impression from Panel I. Additionally, P/R is not significant for $BH(t33)$ in Panel II and results for GW/TA are insignificant for $BH(o10)$ and $BH(t33)$.

The coefficient for $log(MCap)$ is positive for $BH(any)$ but negative for the other specifications. Together with the results from Panel IV (including only non-financial blockholders but also only companies with a market capitalization of above USD 1bn; see table 5.11 in the Appendix), where coefficients are negative in all cases, my impression that it is more likely for small companies (when only looking at companies with a market capitalization of above USD 1bn) to have blockholders than for larger ones is confirmed. Coefficients and z-values however are slightly smaller for non-financial blockholders than for all blockholders. It is unclear whether that is due to the smaller sample size or if there is a causal link (which would mean that for the sample with companies above USD 1bn market capitalization, non-financial blockholders are preferring slightly larger companies than other blockholders).

While results for P/B won't allow any conclusions, there is a positive relationship between P/E and reported blockholdership for all BH variants. Intuitively, I would have expected companies with a lower price-to-earnings ratio to have a higher likelihood of blockholders, as previous results and literature suggests long-term blockholders are preferring lower-valued companies. However, a possible explanation could be that companies with a negative P/E are missing in the sample which could lead to a distortion of results.

Price-to-revenue results for non-financial blockholders are similar to Panel I, confirming that companies with low price-to-revenue ratios are more likely to have reporting non-financial blockholders. However, the result for companies with a cumulative blockholdership of above 33% is insignificant for this Panel. Results for the equity ratio ER confirm blockholders preference for companies with a comparatively high equity ratio, which is again most pronounced for companies with at least one blockholder reporting a stake above 10%.

The dividend-payout ratio DR displays significant negative coefficients for all specifications as it did in Panel I, albeit with slightly lower coefficients and z-values, especially for $BH(o10)$ and $BH(t33)$ (the coefficient for $BH(t33)$ is even slightly positive when only looking at companies above USD 1bn market capitalization in Panel IV). While holders of large blocks may in general be more interested in the long-term health of the company than in short-term cash distributions, non-financial blockholders with large ownership percentages are probably dependent on dividends as well, at least in some cases.

For GW/TA, coefficients don't support the impression from Panel I that companies with a high Goodwill-to-assets ratio are more likely to have reporting blockholders. While the coefficient is still significantly positive for $BH(any)$, results for the other measures are very low and noisy. Further research would be necessary to determine if the relationship of blockholdership and the goodwill-to-assets ratio is indeed different for all blockholders or if the result from Panel I was due to some bias in our sample.

Table 5.9:
Regression Analysis: Determinants of Blockholdership II

This table reports the logistic regression estimates, where the dependent variables are different proxies for Blockholdership BH; only non-financial blockholders are counted. BH is 1 if the respective conditions are met by the company: i) $BH(any)$ is one if the company has at least one blockholder who reported any ownership percentage in the firm-year; ii) $BH(o10)$ is one if at least one blockholder has declared a stake of equal or above 10% and iii) $BH(t33)$ is one if the ownership stake reported by all blockholders together for the firm-year is above 33%. The independent variables are: i) $log(MCap)$, the logarithm of the company's market capitalization; ii) P/B, the price-to-book ratio; iii) P/E, the price-to-earnings ratio; iv) P/R, the price-to-revenues ratio; v) ER, the equity-ratio; vi) DR, the dividend-payout-ratio and vii) GW/TA, the ratio of goodwill to total assets. Additionally, I add Time Fixed Effects (for all years in the sample) and Industry Fixed Effects (for SIC-1). P-values are calculated with a Wald test. Period: November 1993 to May 2021.

Panel II: All companies, Only non-financial Blockholders

	$BH(any)$		$BH(o10)$		$BH(t33)$	
	Coef.	z-val	Coef.	z-val	Coef.	z-val
$Intercept$	−10.0458***	−10.02	−7.7261***	−7.70	−6.8985***	−6.86
$log(MCap)$	0.0789***	24.03	−0.0390***	−10.78	−0.0797***	−17.16
P/B	−0.0090***	−5.69	0.0016	0.96	0.0094***	4.74
P/E	0.0004***	3.41	0.0006***	5.79	0.0005***	4.17
P/R	−0.0027***	−5.47	−0.0015***	−2.68	−0.0008	−1.14
ER	0.1805***	5.44	0.3287***	8.97	0.1345***	2.85
DR	−3.5151***	−15.11	−1.1006***	−4.12	−1.3644***	−3.89
GW/TA	0.8718***	13.97	0.0562	0.88	−0.0379	−0.46
Time & Industry FE	yes		yes		yes	
Pseudo R^2	0.14		0.06		0.05	
Obs	101229		101229		101229	
$\sum(BH = 1)$	55800		23640		11942	

$^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$

5.6 Conclusion and Next Steps

In this paper, I introduce a publicly-available and easily-usable dataset consisting of data extracted from SEC Form 13D, D/A, G and G/A filings. In contrast to Form 13F filings that contain all shareholdings of institutional investment managers, these filings are not required to be filed in a computer-readable format, which makes accessing them difficult and cumbersome.

While some commercial suppliers offer data products containing parts of the Form 13D/G(A) data, this database is to the best of my knowledge the only publicly-available, non-commercial database that contains up-to-date, accurate and complete filing data compiled from these forms.

I hope that using this data enables researchers to get a better understanding of e.g. the reception of new ownership data and information assimilation in financial markets, especially with regards to non-financial blockholders. Combined with other data sources, this data could lead to new insights into the effects of blockholdership and hint at ways to reduce information asymmetry and ultimately improve market efficiency (e.g. through improved, targeted regulation or additional incentives for long-term ownership).

Additionally, the different characteristics of short-term- and long-term-oriented shareholders and their effects on the long-term health of the economy are an interesting and relevant area of research. I plan to use the blockholder data to conduct an empirical study on whether companies with large long-term blockholders outperform companies with a high public float in the long-term using an asset pricing approach.

While the empirical results in chapter 5.5 only scratch the surface of insights contained in the data, they contribute some results to the existing literature on blockholders. I find that the likelihood of a company being the subject of a blockholder filing is higher for companies with medium-sized market capitalization, low price-to-revenue ratios, high equity-ratios and low dividend-payout ratios. Holders of larger ownership percentages prefer companies with even higher equity-ratios than other blockholders and results are generally similar for non-financial blockholders and when only looking at companies above a market capitalization of USD 1bn.

Appendix 5.A - Additional Tables & Figures

Figure 5.8: Number of Filings reported per Year

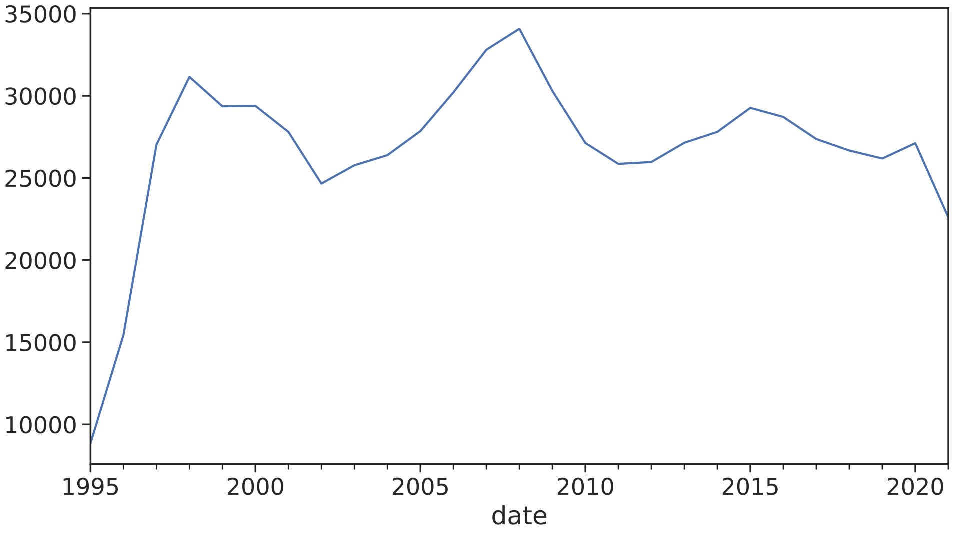


Figure 5.9: Number of Filings reported per Quarter

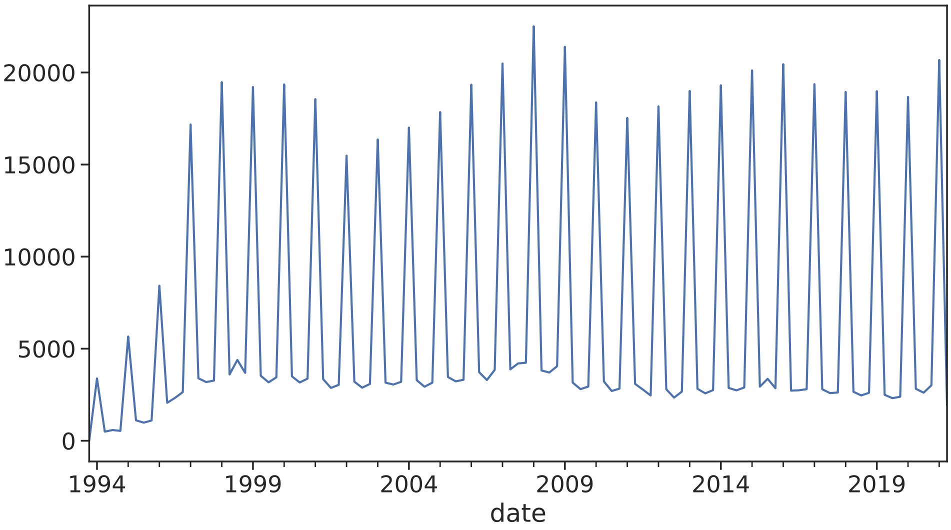


Figure 5.10: Reported Group Membership for Form 13D and 13G (/A) Filings

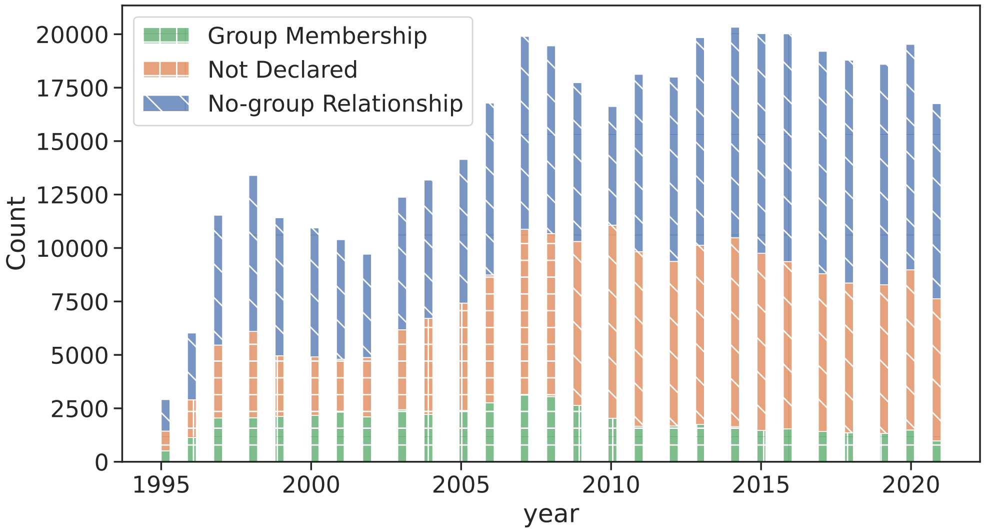


Table 5.10:
Regression Analysis: Determinants of Blockholdership III

This table reports the logistic regression estimates, where the dependent variables are different proxies for Blockholdership BH; companies with a market capitalization of below USD 1bn are dropped from the sample. BH is 1 if the respective conditions are met by the company: i) $BH(any)$ is one if the company has at least one blockholder who reported any ownership percentage in the firm-year; ii) $BH(o10)$ is one if at least one blockholder has declared a stake of equal or above 10% and iii) $BH(t33)$ is one if the ownership stake reported by all blockholders together for the firm-year is above 33%. The independent variables are: i) $log(MCap)$, the logarithm of the company's market capitalization; ii) P/B, the price-to-book ratio; iii) P/E, the price-to-earnings ratio; iv) P/R, the price-to-revenues ratio; v) ER, the equity-ratio; vi) DR, the dividend-payout-ratio and vii) GW/TA, the ratio of goodwill to total assets. Additionally, I add Time Fixed Effects (for all years in the sample) and Industry Fixed Effects (for SIC-1). P-values are calculated with a Wald test. Period: November 1993 to May 2021.

Panel III: Only companies above USD 1bn Marketcap, all Blockholders

	$BH(any)$		$BH(o10)$		$BH(t33)$	
	Coef.	z-val	Coef.	z-val	Coef.	z-val
$Intercept$	0.7961	0.77	3.3809***	3.29	4.3127***	4.18
$log(MCap)$	−0.3374***	−31.93	−0.4561***	−45.41	−0.5039***	−44.45
P/B	0.0311***	8.75	0.0227***	9.47	0.0186***	7.66
P/E	−0.0007***	−3.19	0.0009***	4.91	0.0012***	6.14
P/R	−0.0047***	−3.65	−0.0033**	−2.53	−0.0032**	−2.28
ER	0.2863***	3.97	0.5924***	10.02	0.3913***	6.22
DR	−10.8118***	−22.88	−6.7899***	−14.81	−6.9916***	−13.77
GW/TA	1.1801***	10.40	0.4605***	5.42	0.3004***	3.35
Time & Industry FE	yes		yes		yes	
Pseudo R^2	0.15		0.11		0.12	
Obs	38575		38575		38575	
$\sum(BH = 1)$	29617		16570		11938	

$^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$

Table 5.11:
Regression Analysis: Determinants of Blockholdership IV

This table reports the logistic regression estimates, where the dependent variables are different proxies for Blockholdership BH; companies with a market capitalization of below USD 1bn are dropped from the sample and only non-financial blockholders are counted. BH is 1 if the respective conditions are met by the company: i) $BH(any)$ is one if the company has at least one blockholder who reported any ownership percentage in the firm-year; ii) $BH(o10)$ is one if at least one blockholder has declared a stake of equal or above 10% and iii) $BH(t33)$ is one if the ownership stake reported by all blockholders together for the firm-year is above 33%. The independent variables are: i) $log(MCap)$, the logarithm of the company's market capitalization; ii) P/B, the price-to-book ratio; iii) P/E, the price-to-earnings ratio; iv) P/R, the price-to-revenues ratio; v) ER, the equity-ratio; vi) DR, the dividend-payout-ratio and vii) GW/TA, the ratio of goodwill to total assets. Additionally, I add Time Fixed Effects (for all years in the sample) and Industry Fixed Effects (for SIC-1). P-values are calculated with a Wald test. Period: November 1993 to May 2021.

Panel IV: Only companies above USD 1bn Marketcap, only non-financial Blockholders

	$BH(any)$		$BH(o10)$		$BH(t33)$	
	Coef.	z-val	Coef.	z-val	Coef.	z-val
$Intercept$	0.0809	0.08	3.1359***	3.03	−0.1300	−0.12
$log(MCap)$	−0.3017***	−30.92	−0.4414***	−36.87	−0.2946***	−18.61
P/B	0.0145***	5.35	0.0120***	4.77	0.0207***	7.04
P/E	0.0003	1.21	0.0007***	3.86	0.0010***	4.06
P/R	−0.0038***	−3.09	−0.0013	−1.02	0.0006	0.45
ER	0.1773***	2.78	0.3017***	4.56	0.1552*	1.73
DR	−9.4932***	−20.45	−3.0318***	−5.96	0.2045	0.32
GW/TA	0.5058***	5.39	−0.3730***	−3.84	−0.1826	−1.41
Time & Industry FE	yes		yes		yes	
Pseudo R^2	0.16		0.09		0.04	
Obs	38575		38575		38575	
$\sum(BH = 1)$	24534		9347		3988	

$^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$

Appendix 5.B - Technical Details on Blockholder Parsing

All in all, the parser and downloader used for creating the database contain more than 1,500 lines of code. It is planned to open-source the code in order to enable other researchers to adapt it to other semi-structured filing types as well. Below, the code for parsing the ownership percentages (as discussed in chapter5.4.1) is given.

Percentage Parsing Code

```
def parse_holding_percentage(self, document_text):
    percentage_lines=self.get_percentage_lines(document_text)
    percentages=self.get_percentages(percentage_lines)
    return self.get_max_percentage(percentages)

def get_percentage_lines(self, body):
    body = body.lower()
    split_lines = body.split("\n")
    split_lines = list(filter(None, split_lines))
    split_lines = list(filter(str.strip, split_lines))

    split_lines = split_lines[0:2999]

    # find rows containing "percent"
    percent_idx = [i for i, item in enumerate(split_lines) if re.search('percent', item)]
    # create list with all possible lines with percentage by index location
    possible_lines = []
    for line in percent_idx:
        possible_lines += list(range(line, line + 10))

    final_lines = []
    for i, line in enumerate(split_lines):
        if i in possible_lines:
            final_lines.append(line)
    return (final_lines)

def get_percentages(self, lines, mode='standard'):
    if mode == 'standard':
        # non decimal or start, then 1-3 decimals, dot or comma with 1 or 2 decimals or nothing,
        whitespace, percentage
        re_percentage = re.compile(’(^|\s|(\.\.)|:|=)\d{1,3}([\.\,]\d{1,3})?\s*(?=%)’)
    else:
        # search vor percentages without percentage sign when nothing found
        re_percentage = re.compile(’(^|\s|(\.\.)|:|=)\d{1,3}[\.\,]\d{1,3}($|\s)’)
    matches = []
    for line in lines:
        for match in re_percentage.finditer(line):
            match_processed = match.group().replace(’,’, ’.’).replace(’ ’, ’’)
            .replace(’:’, ’’).replace(’=’,’’)
            if match_processed.startswith(’..’):
                match_processed = match_processed[2:]
            matches.append(match_processed)
    if ((len(matches) == 0) or (
            matches == ["5"])) and mode == ’standard’:  # try again when we did not find anything
        return self.get_percentages(lines, mode=’no_percentage_sign’)
    else:
        # returns list of found percentage strings
        return matches

def get_max_percentage(self, percentages):
    percentages = np.array([float(x) for x in percentages if float(x) <= 100.0])
    if percentages.size > 0:
        percentages = self.check_for_int_floats(percentages)
        return choose_most_often(percentages)
    else:
        return np.nan

    def check_for_int_floats(self, percentages):
    integers = np.zeros(len(percentages))
    for i, x in enumerate(percentages):
        if x == int(x):
            integers[i] = 1
    if integers.min() == 0:
        return percentages[integers == 0]
    else:
        if percentages.size > 1:
            # filter out known error sizes if multiple hits found
            new_percentages = np.array([x for x in percentages if x not in (5, 9, 10, 11)])
            if new_percentages.size > 0:
                return new_percentages
        return percentages

def choose_most_often(item_list, tie=’max’):
if len(item_list) > 1:
    counts = Counter(item_list).most_common()
    if len(counts)>1 and counts[0][1] == counts[1][1]:
        if tie==’max’:
            return max([counts[0][0], counts[1][0]])
        else:
            return counts[0][0]
    else:
        return counts[0][0]
else:
    return item_list[0]
```

Appendix 5.C - Description of Columns in the Blockholder Database

- **cik**: CIK number in the header of the filing
- **name**: Company name in the header of the filing
- **type**: Filing Type (can be 'SC 13D/A', 'SC 13G/A', 'SC 13G', 'SC 13D') in the header of the filing
- **date**: Date in the header of the filing
- **link**: Permanent link to the filing on SEC EDGAR
- **file**: Filename of the Filing
- **acceptance_dt**: Datetime the filing was received by the SEC
- **accession_number**: Unique number of the filing
- **conformed_type**: Self-reported type of filing. Normally, "type" should be used instead.
- **document_count**: Number of documents contained in the filings.
- **filed_as_of**: Can be a future date if filings has been pre-filed in some rare cases
- **changed_as_of**: Date of changes in the filing
- **group_members**: List of Group Members, divided by ";"
- **text-link**: Link to the filing in text-only form
- **sequence**: Sequence number (for filings with multiple documents)
- **document_type**: Type of document for filings containing multiple documents
- **filename**: Name of the text file
- **description**: Free-form text with a description of the filing (e.g. "SCHEDULE 13 G - AMENDMENT #2)

- **cusip**: CUSIP Number of the subject company, parsed from the filing
- **percent_held**: Ownership percentage, parsed from the filing (in case multiple ownership percentages are reported, highest number is taken)
- **group_membership**: Declaration of Group Membership (can be "a" for declared group membership, "b" for declared no-group association, "False" for no declaration)
- **investor_type**: Self-declared type of investor, can be 'individual', 'employee_benefit', 'holding_company', 'corporation', 'investment_company', 'investment_advisor', 'insurance', 'broker', 'bank', 'partnership', 'other', 'savings_association', 'church' or 'non_us_institution'
- **below5**: True if the filing reports that ownership has ceased to be above 5%, False otherwise
- **source_of_funds**: self-declared source of funds (13D only), can be affiliate', 'bank', 'working_capital', 'personal_funds', 'other', 'subject_company'
- **legal_proceedings**: True if disclosure of legal proceedings is required for the filer, False otherwise
- **previous_13g**: True if a previous Form 13G filing is declared, False otherwise
- **subj_company_name**: The name of the subject company
- **subj_cik**: The CIK of the subject company
- **subj_sic**: The SIC classification of the subject company (e.g. "TELEVISION BROADCASTING STATIONS")
- **subj_sic_code**: The SIC classification code of the subject company
- **subj_irs**: The IRS number of the subject company
- **subj_state_of_incorporation**: The state of incorporation of the subject company
- **subj_fiscal_year_end**: The fiscal year-end of the subject company
- **subj_form_type**: The filing type for the subject company

- **subj_sec_act**: The relevant SEC act for the subject company (e.g. "1934 Act")
- **subj_sec_file_number**: The SEC file number of the subject company
- **subj_film_number**: The film number of the subject company
- **subj_street1**: The first line of the street address of the subject company
- **subj_street2**: The second line of the street address of the subject company
- **subj_city**: The city of the subject company
- **subj_state**: The state of the subject company
- **subj_zip**: The ZIP code of the subject company
- **subj_phone**: The phone number of the subject company
- **subj_mail_street1**: The first line of the mailing address of the subject company
- **subj_mail_street2**: The second line of the mailing address of the subject company
- **subj_mail_city**: The city of the mailing address of the subject company
- **subj_mail_state**: The state of the mailing address of the subject company
- **subj_mail_zip**: The ZIP code of the mailing address of the subject company
- **subj_former_name**: The former name of the subject company
- **subj_date_name_change**: The date of the name change of the subject company
- **fil_company_name**: The name of the filing entity
- **fil_cik**: The CIK of the filing entity
- **fil_sic**: The SIC classification of the filing entity
- **fil_sic_code**: The SIC classification code of the filing entity
- **fil_irs**: The IRS number of the filing entity

- **fil_state_of_incorporation**: The state where the filing entity is incorporated
- **fil_fiscal_year_end**: The fiscal year end of the filing entity
- **fil_form_type**: The filing type as declared by the filing entity
- **fil_street1**: The first line of the address of the filing entity
- **fil_street2**: The second line of the address of the filing entity
- **fil_city**: The city of the filing entity
- **fil_state**: The state of the filing entity
- **fil_zip**: The ZIP code of the filing entity
- **fil_phone**: The phone number of the filing entity
- **rule13d**: True is the filing is pursuant to Rule 13-D, False otherwise
- **fil_mail_street1**: The first line of the mailing address of the filing entity
- **fil_mail_street2**: The second line of the mailing address of the filing entity
- **fil_mail_city**: The city of the mailing address of the filing entity
- **fil_mail_state**: The state of the mailing address of the filing entity
- **fil_mail_zip**: The ZIP code of the mailing address of the filing entity
- **fil_former_name**: The former name of the filing entity
- **fil_date_name_change**: The date of the name change of the filing entity
- **subj_former_name_2**: The second former name of the subject company
- **subj_date_name_change_2**: The date of the second name change of the subject company
- **fil_former_name_2**: The second former name of the filing entity
- **fil_date_name_change_2**: The second date of the name change of the filing entity

A detailed description of the SEC EDGAR Public Dissemination Service Technical Specification can be found at: `https://www.sec.gov/info/edgar/specifications/pds_dissemination_spec.pdf`.

Bibliography

Ahmad, K., Han, J., Hutson, E., Kearney, C., and Liu, S. (2016). Media-expressed negative tone and firm-level stock returns. Journal of Corporate Finance, 37:152–172.

Alford, A. W., Jones, J. J., and Zmijewski, M. E. (1994). Extensions and violations of the statutory SEC form 10-K filing requirements. Journal of Accounting and Economics, 17(1-2):229–254.

Almadi, H., Rapach, D. E., and Suri, A. (2014). Return Predictability and Dynamic Asset Allocation: How Often Should Investors Rebalance? The Journal of Portfolio Management, 40(4):16–27.

Aminadav, G. and Papaioannou, E. (2020). Corporate Control around the World. The Journal of Finance, 75(3):1191–1246.

Ammann, M., Moellenbeck, M., and Schmid, M. M. (2011). Feasible momentum strategies in the US stock market. Journal of Asset Management, 11(6):362–374.

Andres, C. (2008). Large shareholders and firm performance—An empirical examination of founding-family ownership. Journal of Corporate Finance, 14(4):431–445.

Angel, J. (2021). Gamestonk: What Happened and What to Do about It. SSRN Scholarly Paper ID 3782195.

Anton, M., Gine, M., and Schmalz, M. C. (2016). Common Ownership, Competition, and Top Management Incentives. SSRN Scholarly Paper ID 2802332.

Asness, C. and Frazzini, A. (2013). The Devil in HML's Details. The Journal of Portfolio Management, 39(4):49–68.

Bachelier, L. (1900). Théorie de la spéculation. Annales scientifiques de l'École Normale Supérieure, 17:21–86.

Backus, M., Conlon, C., and Sinkinson, M. (2019). Common Ownership in America: 1980-2017. Technical Report w25454, National Bureau of Economic Research.

Baker, M. and Wurgler, J. (2007). Investor Sentiment in the Stock Market. Journal of Economic Perspectives, 21(2):129–152.

Bali, T. G., Cakici, N., and Whitelaw, R. F. (2011). Maxing out: Stocks as lotteries and the cross-section of expected returns. Journal of Financial Economics, 99(2):427–446.

Bali, T. G., Engle, R. F., and Murray, S. (2016). Empirical Asset Pricing: The Cross Section of Stock Returns. John Wiley & Sons.

Barber, B. M., Odean, T., and Zhu, N. (2006). Do Noise Traders Move Markets? SSRN Scholarly Paper ID 869827.

Barberis, N. and Thaler, R. (2003). A survey of behavioral finance. In Handbook of the Economics of Finance, volume 1 of Financial Markets and Asset Pricing, pages 1053–1128. Elsevier.

Barillas, F. and Shanken, J. (2017). Which Alpha? The Review of Financial Studies, 30(4):1316–1338.

Barrot, J.-N., Kaniel, R., and Sraer, D. (2016). Are retail traders compensated for providing liquidity? Journal of Financial Economics, 120(1):146–168.

Baskin, J. B. (1988). The Development of Corporate Financial Markets in Britain and the United States, 1600-1914: Overcoming Asymmetric Information. The Business History Review, 62(2):199–237.

Basu, S. (1983). The relationship between earnings' yield, market value and return for NYSE common stocks: Further evidence. Journal of Financial Economics, 12(1):129–156.

Battalio, R., Hatch, B., and Jennings, R. (2004). Toward a National Market System for U.S. Exchange–listed Equity Options. The Journal of Finance, 59(2):933–962.

Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., and Blackburn, J. (2020). The Pushshift Reddit Dataset. Proceedings of the International AAAI Conference on Web and Social Media, 14:830–839.

Behrendt, S. and Schmidt, A. (2018). The Twitter myth revisited: Intraday investor sentiment, Twitter activity and individual-level stock return volatility. Journal of Banking & Finance, 96:355–367.

Bender, J. and Wang, T. (2016). Can the Whole Be More Than the Sum of the Parts? Bottom-Up versus Top-Down Multifactor Portfolio Construction. The Journal of Portfolio Management, 42(5):39–50.

Black, F. (1972). Capital Market Equilibrium with Restricted Borrowing. The Journal of Business, 45(3):444–55.

Block, S. B. and French, D. W. (2002). The effect of portfolio weighting on investment performance evaluation: The case of actively managed mutual funds. Journal of Economics and Finance, 26(1):16–30.

Boehmer, E., Jones, C. M., and Zhang, X. (2021). Tracking Retail Investor Activity. Journal of Finance, forthcoming.

Bondt, W. F. M. D. and Thaler, R. (1985). Does the Stock Market Overreact? The Journal of Finance, 40(3):793–805.

Bradley, D., Hanousek Jr., J., Jame, R., and Xiao, Z. (2021). Place Your Bets? The Market Consequences of Investment Advice on Reddit's Wallstreetbets. SSRN Scholarly Paper ID 3806065.

Brennan, M. J. and Wang, A. (2007). Asset Pricing and Mispricing. SSRN Scholarly Paper ID 912814.

Broadstock, D. C. and Zhang, D. (2019). Social-media and intraday stock returns: The pricing power of sentiment. Finance Research Letters, 30:116–123.

Bushee, B. J. (2001). Do Institutional Investors Prefer Near-Term Earnings over Long-Run Value?*. Contemporary Accounting Research, 18(2):207–246.

Campbell, G., Turner, J. D., and Walker, C. B. (2012). The role of the media in a bubble. Explorations in Economic History, 49(4):461–481.

Carhart, M. M. (1997). On Persistence in Mutual Fund Performance. The Journal of Finance, 52(1):57–82.

Chan, L. K. C., Hamao, Y., and Lakonishok, J. (1991). Fundamentals and Stock Returns in Japan. The Journal of Finance, 46(5):1739–1764.

Clifford, C. (2008). Value creation or destruction? hedge funds as shareholder activists. Journal of Corporate Finance, 14(4):323–336.

Clifford, C. P. and Lindsey, L. (2016). Blockholder Heterogeneity, CEO Compensation, and Firm Performance. Journal of Financial and Quantitative Analysis, 51(5):1491–1520.

Cochrane, J. (1999). Portfolio Advice for a Multifactor World. Technical Report w7170, National Bureau of Economic Research.

Cochrane, J. H. (1996). A Cross-Sectional Test of an Investment-Based Asset Pricing Model. Journal of Political Economy, 104(3):572–621.

Cornell, B. and Damodaran, A. (2014). Tesla: Anatomy of a Run-Up Value Creation or Investor Sentiment? SSRN Scholarly Paper ID 2429778.

Cremers, M. (2013). Should Benchmark Indices Have Alpha? Revisiting Performance Evaluation. Critical Finance Review, 2(1):1–48.

Cronqvist, H. and Fahlenbrach, R. (2009). Large Shareholders and Corporate Policies. The Review of Financial Studies, 22(10):3941–3976.

Daniel, K., Hirshleifer, D., and Teoh, S. H. (2002). Investor psychology in capital markets: evidence and policy implications. Journal of Monetary Economics, 49(1):139–209.

Dlugosz, J., Fahlenbrach, R., Gombers, P., and Metrick, A. (2006). Large blocks of stock: Prevalence, size, and measurement. Journal of Corporate Finance, 12(3):594–618.

Dyl, E. A. and Maberly, E. D. (1992). Odd-Lot Transactions around the Turn of the Year and the January Effect. The Journal of Financial and Quantitative Analysis, 27(4):591–604.

Easton, P. D. and Zmijewski, M. E. (1993). SEC Form 10K/10Q Reports and Annual Reports to Shareholders: Reporting Lags and Squared Market Model Prediction Errors. Journal of Accounting Research, 31(1):113–129.

Edmans, A. (2009). Blockholder Trading, Market Efficiency, and Managerial Myopia. The Journal of Finance, 64(6):2481–2513.

Edmans, A. and Holderness, C. G. (2017). Chapter 8 - Blockholders: A Survey of Theory and Evidence. In Hermalin, B. E. and Weisbach, M. S., editors, The

Handbook of the Economics of Corporate Governance, volume 1 of The Handbook of the Economics of Corporate Governance, pages 541–636. North-Holland.

Fama, E. F. (1965). The Behavior of Stock-Market Prices. The Journal of Business, 38(1):34–105.

Fama, E. F. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. The Journal of Finance, 25(2):383–417.

Fama, E. F. (1991). Efficient Capital Markets: II. The Journal of Finance, 46(5):1575–1617.

Fama, E. F. and French, K. R. (1992). The Cross-Section of Expected Stock Returns. The Journal of Finance, 47(2):427–465.

Fama, E. F. and French, K. R. (1993). Common risk factors in the returns on stocks and bonds. Journal of Financial Economics, 33(1):3–56.

Fama, E. F. and French, K. R. (1995). Size and Book-to-Market Factors in Earnings and Returns. The Journal of Finance, 50(1):131–155.

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

Fama, E. F. and French, K. R. (2018). Choosing factors. Journal of Financial Economics, 128(2):234–252.

Fama, E. F. and MacBeth, J. D. (1973). Risk, Return, and Equilibrium: Empirical Tests. Journal of Political Economy, 81(3):607–636.

Farrell, M., Green, T. C., Jame, R., and Markov, S. (2020). The Democratization of Investment Research and the Informativeness of Retail Investor Trading. SSRN Scholarly Paper ID 3222841.

Gao, L., Han, Y., Zhengzi Li, S., and Zhou, G. (2018). Market intraday momentum. Journal of Financial Economics, 129(2):394–414.

García, D. (2013). Sentiment during Recessions. The Journal of Finance, 68(3):1267–1300.

Gelman, A. and 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. Working Paper Series, Department of Statistics, Columbia University.

Gibbons, M. R., Ross, S. A., and Shanken, J. (1989). A test of the efficiency of a given portfolio. Econometrica, 57(5):1121–1152.

Gloßner, S. (2019). Investor Horizons, Long-Term Blockholders, and Corporate Social Responsibility. Journal of Banking & Finance, 103:78–97.

Goyal, A. (2019). Which Factors? Editorial Commentary. Review of Finance, 23(1).

Hadlock, C. J. and Schwartz-Ziv, M. (2018). Blockholder Heterogeneity, Multiple Blocks, and the Dance Between Blockholders. SSRN Scholarly Paper ID 2894688.

Han, B. and Kumar, A. (2013). Speculative Retail Trading and Asset Prices. The Journal of Financial and Quantitative Analysis, 48(2):377–404.

Hanauer, M. X. (2020). A Comparison of Global Factor Models. SSRN Scholarly Paper ID 3546295.

Harvey, C. R. (2017). Presidential Address: The Scientific Outlook in Financial Economics. The Journal of Finance, 72(4):1399–1440.

Harvey, C. R. and Liu, Y. (2015). Backtesting. The Journal of Portfolio Management, 42(1):13–28.

Harvey, C. R., Liu, Y., and Zhu, H. (2016). ... and the Cross-Section of Expected Returns. The Review of Financial Studies, 29(1):5–68.

Hasso, T., Müller, D., Pelster, M., and Warkulat, S. (2021). Who Participated in the GameStop Frenzy? Evidence from Brokerage Accounts. SSRN Scholarly Paper ID 3792095.

Hayek, F. A. (1945). The Use of Knowledge in Society. The American Economic Review, 35(4):519–530.

He, H. and Modest, D. M. (1995). Market Frictions and Consumption-Based Asset Pricing. Journal of Political Economy, 103(1):94–117.

Holderness, C. G. (2003). A survey of blockholders and corporate control. Economic Policy Review, 9(Apr):51–64.

Holderness, C. G. (2009). The Myth of Diffuse Ownership in the United States. The Review of Financial Studies, 22(4):1377–1408.

Hou, K., Mo, H., Xue, C., and Zhang, L. (2019). Which Factors? Review of Finance, 23(1):1–35.

Hou, K., Xue, C., and Zhang, L. (2015). Digesting Anomalies: An Investment Approach. The Review of Financial Studies, 28(3):650–705.

Hou, K., Xue, C., and Zhang, L. (2020). Replicating Anomalies. The Review of Financial Studies, 33(5):2019–2133.

Hsu, P.-H. and Hsu, Y.-C. (2006). A Stepwise Spa Test for Data Snooping and its Application on Fund Performance Evaluation. SSRN Scholarly Paper ID 885364.

Huntington-Klein, N., Arenas, A., Beam, E., Bertoni, M., Bloem, J. R., Burli, P., Chen, N., Grieco, P., Ekpe, G., Pugatch, T., Saavedra, M., and Stopnitzky, Y.

(2021). The influence of hidden researcher decisions in applied microeconomics. Economic Inquiry, 59(3):944–960.

Ioannidis, J. P. A., Stanley, T. D., and Doucouliagos, H. (2017). The Power of Bias in Economics Research. The Economic Journal, 127(605):F236–F265.

Jones, C. M., Reed, A. V., and Waller, W. (2021). When Brokerages Restrict Retail Investors, Does the Game Stop? SSRN Scholarly Paper ID 3804446.

Kan, R., Wang, X., and Zheng, X. (2019). In-Sample and Out-of-Sample Sharpe Ratios of Multi-Factor Asset Pricing Models. SSRN Scholarly Paper ID 3454628.

Kaniel, R., Saar, G., and Titman, S. (2008). Individual Investor Trading and Stock Returns. The Journal of Finance, 63(1):273–310.

Kessler, S., Scherer, B., and Harries, J. (2020). Value by Design? Journal of Portfolio Management, 46(5):25–43.

Kumar, A. and Lee, C. M. C. (2006). Retail Investor Sentiment and Return Comovements. The Journal of Finance, 61(5):2451–2486.

Kumar, A., Ruenzi, S., and Ungeheuer, M. (2020). Daily Winners and Losers. SSRN Scholarly Paper ID 2931545.

Lakonishok, J., Shleifer, A., and Vishny, R. W. (1994). Contrarian Investment, Extrapolation, and Risk. The Journal of Finance, 49(5):1541–1578.

Ledoit, O. and Wolf, M. (2008). Robust performance hypothesis testing with the Sharpe ratio. Journal of Empirical Finance, 15(5):850–859.

Lintner, J. (1965). The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets. The Review of Economics and Statistics, 47(1):13–37.

Long, C., Lucey, B. M., and Yarovaya, L. (2021). "I Just Like the Stock" versus "Fear and Loathing on Main Street" : The Role of Reddit Sentiment in the GameStop Short Squeeze by Cheng Long, Brian M. Lucey, Larisa Yarovaya :: SSRN. SSRN Scholarly Paper ID 3822315.

Luttmer, E. (1996). Asset Pricing in Economies with Frictions. Econometrica, 64(6):1439–67.

Lynch, A. W. and Balduzzi, P. (2000). Predictability and Transaction Costs: The Impact on Rebalancing Rules and Behavior. The Journal of Finance, 55(5):2285–2309.

Mandelbrot, B. B. (1963). The variation of certain speculative prices. The Journal of Business, 36(4):394–419.

Mclean, R. D. and Pontiff, J. (2016). Does Academic Research Destroy Stock Return Predictability? The Journal of Finance, 71(1):5–32.

Mehran, H. (1995). Executive compensation structure, ownership, and firm performance. Journal of Financial Economics, 38(2):163–184.

Meredith, M. (2007). A librarian's guide to the securities and exchange commission's filings. The Reference Librarian, 48(1):35–55.

Newey, W. K. and West, K. D. (1987). A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. Econometrica, 55(3):703–708.

Nisar, T. M. and Yeung, M. (2018). Twitter as a tool for forecasting stock market movements: A short-window event study. The Journal of Finance and Data Science, 4(2):101–119.

O'Hara, M., Yao, C., and Ye, M. (2014). What's Not There: Odd Lots and Market Data. The Journal of Finance, 69(5):2199–2236.

Ozik, G., Sadka, R., and Shen, S. (2021). Flattening the Illiquidity Curve: Retail Trading During the COVID-19 Lockdown. SSRN Scholarly Paper ID 3663970.

Pedersen, L. H. (2021). Game On: Social Networks and Markets. SSRN Scholarly Paper ID 3794616.

Plyakha, Y., Uppal, R., and Vilkov, G. (2014). Equal or Value Weighting? Implications for Asset-Pricing Tests. SSRN Scholarly Paper ID 1787045.

Roll, R. (1977). A critique of the asset pricing theory's tests Part I: On past and potential testability of the theory. Journal of Financial Economics, 4(2):129–176.

Rosenberg, B., Reid, K., and Lanstein, R. (1985). Persuasive evidence of market inefficiency. The Journal of Portfolio Management, 11(3):9–16.

Samuelson, P. A. (1965). Proof-That-Properly-Anticipated-Prices-Fluctuate-Randomly-Paul-A.-Samuelson-1965.pdf. Industrial Management Review, 6(2):41–49.

Schwartz-Ziv, M. and Volkova, E. (2020). Is Blockholder Diversity Detrimental? SSRN Scholarly Paper ID 3621939.

Sharpe, W. (1964). CAPITAL ASSET PRICES: A THEORY OF MARKET EQUILIBRIUM UNDER CONDITIONS OF RISK. Journal of Finance, 19(3):425–442.

Shleifer, A. (2000). Inefficient Markets: An Introduction to Behavioural Finance. Oxford University Press.

Shleifer, A. and Vishny, R. W. (1986). Large Shareholders and Corporate Control. Journal of Political Economy, 94(3):461–488.

Shleifer, A. and Vishny, R. W. (1997). The Limits of Arbitrage. The Journal of Finance, 52(1):35–55.

Smith, D. M. and Desormeau, W. (2006). Optimal Rebalancing Frequency for Stock-Bond Portfolios. SSRN Scholarly Paper ID 2458618.

Stambaugh, R. F. and Yuan, Y. (2017). Mispricing Factors. The Review of Financial Studies, 30(4):1270–1315.

Sun, L., Najand, M., and Shen, J. (2016). Stock return predictability and investor sentiment: A high-frequency perspective. Journal of Banking & Finance, 73:147–164.

Tetlock, P. C. (2007). Giving Content to Investor Sentiment: The Role of Media in the Stock Market. The Journal of Finance, 62(3):1139–1168.

Umar, Z., Gubareva, M., Yousaf, I., and Ali, S. (2021). A tale of company fundamentals vs sentiment driven pricing: The case of GameStop. Journal of Behavioral and Experimental Finance, 30:100501.

van der Beck, P. and Jaunin, C. (2021). The Equity Market Implications of the Retail Investment Boom. SSRN Scholarly Paper ID 3776421.

Vasileiou, E., Bartzou, E., and Tzanakis, P. (2021). Explaining Gamestop Short Squeeze using Intraday Data and Google Searches. SSRN Scholarly Paper ID 3805630.

Yan, X. S. and Zheng, L. (2017). Fundamental Analysis and the Cross-Section of Stock Returns: A Data-Mining Approach. The Review of Financial Studies, 30(4):1382–1423.