Peer-reviewed theory does not help predict the cross-section of stock returns

Andrew Y. Chen[1], Alejandro Lopez-Lira[2], and Tom Zimmermann[3]

[1]Federal Reserve Board

[2]University of Florida

[3]University of Cologne

March 7, 2023

Abstract

To examine whether theory helps predict the cross-section of returns, we combine text analysis of publications with out-of-sample tests. Based on the original texts, only 18% of predictors are attributed to risk-based theory. 59% are attributed to mispricing, and 23% have uncertain origins. Post-publication, risk-based predictability decays by 65% (p-value $< 0.1\%$), compared to 50% for non-risk predictors. Out-of-sample, risk-based predictors fail to outperform data-mined accounting predictors that are matched on in-sample summary statistics. Published and data-mined returns rise before in-sample periods end and fall out-of-sample at similar rates. Overall, peer-reviewed research adds little information about future mean returns above naive backtesting.

E-mails: andrew.y.chen@frb.gov, Alejandro.Lopez-Lira@warrington.ufl.edu, tom.zimmermann@uni-koeln.de. Earlier versions of this paper relied on data provided by Sterling Yan and Lingling Zheng, to whom we are grateful. We thank Alec Erb for excellent research assistance. For helpful comments, we thank Albert Menkveld, Ben Knox, Emilio Osambela, Dino Palazzo, Lingling Zheng, and seminar participants at the Federal Reserve Board and the University of Wisconsin-Milwaukee. The views in this paper are not necessarily those of the Federal Reserve Board or the Federal Reserve System.

1 Introduction

Since the creation of the Sharpe-Lintner-Treynor-Mossin CAPM (1962-1966), generations of asset pricing theorists have used risk to understand the cross-section of expected stock returns (Cochrane (2009); Back (2010)). In parallel, empiricists have documented more than 200 determinants of this cross-section (Chen and Zimmermann (2022a)). In this paper, we ask: does the theory help us understand the data?

To answer this question, we combine text analysis of the original papers with out-of-sample tests. Each empirical predictor is published along with text that attempts to explain the origin of return predictability. This text takes on the wisdom of the entire finance profession through the peer review process. In an ideal world, this process picks out the best explanation for each predictor, whether it is risk, mispricing, or some other mechanism. And if risk-based theory helps us understand expected returns, then risk-based predictability should persist out-of-sample. At the very least, predictability that peer review attributes to risk should be more persistent than predictability derived from pure data mining.

We find that peer review attributes only a small minority of predictors to risk. We read the papers corresponding to 191 published predictors in the Chen and Zimmermann (2022a) dataset and assign each predictor to "risk," "mispricing," or "agnostic" based on the arguments made in the texts. Peer review attributes only 18% of predictors to risk. 59% are attributed to mispricing, and 23% have uncertain origins. We validate these results by using software to count the ratio of risk-related words to mispricing-related words. For the median predictor, mispricing-related words are three times more common than risk-related words.

It may be that risk-based theory was slow to develop relative to empirical progress. The data suggest this is true: 25% of predictors published 2005-2016 are attributed to risk, compared to only 6% published 1981-2004. However, peer review can make errors for long stretches of time (Kuhn (1962)). To move past a false paradigm, it is likely that powerful evidence is required (Akerlof and Michaillat (2018)).

In our view, the most powerful evidence is out-of-sample predictability. If a published theory is true, then the results of its empirical tests should continue to hold in the years after the original samples end. Only this kind of test can avoid post-hoc theorizing, which can, in principle, be used to "explain" nearly any empirical pattern (Sonnenschein (1972); Mas-Colell (1977)).

Unfortunately, we find that risk-based predictability largely vanishes out-of-sample. For risk-based predictors, long-short returns formed following the instructions in the original papers decay by 65%

post-publication. Though the sample of risk predictors is relatively small, we can reject the hypothesis of no decay at the 0.1% level. For comparison, mispricing and agnostic predictors decay by 50%. Regression results confirm that having a risk-based explanation is a slightly *negative* signal about external validity. If peer review attributes a predictor to risk, then its mean return going forward is about 35% *lower*.

These results are exactly the opposite of what is implied by asset pricing theory. Risk-based predictability is founded on the concept of equilibrium, implying mean returns that are stable and continue out-of-sample. If anything, the publication of a new risk theory should lead to higher mean returns, as academics teach investors about new risks to avoid. In contrast, mispricing-based predictability should decline out-of-sample as investors learn and push markets toward a more stable equilibrium. Thus, our results imply that peer review either mislabels mispricing as risk or finds unstable risk factors that systematically disappear over time.

This evidence shows that risk-based theory does not prevent out-of-sample decay. But is risk-based theory helpful compared to having no theory at all?

To address this question, we compare the out-of-sample returns of risk-based theory to those from naive data mining. Inspired by Yan and Zheng (2017), we construct 29,000 trading strategies formed by sorting stocks on simple functions of 242 accounting variables. These strategies are formed by (1) dividing one variable by another or (2) taking first differences and then dividing, where the denominators are restricted to be variables that are positive for more than 25% of firms in 1963. Despite the complete lack of economics in this data mining exercise, we find this procedure generates substantial "out-of-sample" returns. Sorting strategies into quintiles based on past returns each year and then trading the bottom quintile earns 50 bps per month, implying thousands of strategies with meaningful out-of-sample returns.

We then match data-mined predictors to published predictors based on their mean returns and t-stats, using the published sample periods. Matching these statistics is important as they determine out-of-sample returns in multiple testing frameworks (Chen and Zimmermann (2020)). Matching the sample periods is important too, as predictability for published predictors is weaker post-2004 (Chordia et al. (2014); Chen and Velikov (2022)). Indeed, we document a similar post-2004 decay in data-mined predictability.

We find that risk-based predictors fail to outperform naive data mining. Risk-based predictability is slightly stronger in the first several years out-of-sample, but performance drops precipitously afterward. For years 10 through 20 out-of-sample, risk-based returns average near zero, while data-mined returns

hover around 25 percent of their in-sample means. These shocking results mean that using risk-based theory is no better than using no theory at all for predicting the cross-section of stock returns.

These results are shocking because economic theory is widely considered the best hope for protecting against data mining bias (Cochrane (2009); Harvey et al. (2016); Harvey (2017)).[1] This bias comes from selecting for the best results from a large set (Chen and Zimmermann (2022b)). Economic theory should restrict this set in a way that tilts toward true predictability and therefore limit the impact of selection bias. Peer review further restricts the set of theories, as only the best theories should make it into the top finance journals.

However, the set of predictors allowed by peer-reviewed theory is in practice quite large. Since Merton (1973), finance researchers have recognized that any proxy for unobserved state variables relevant to the marginal utility of a marginal investor is, in principle, a valid asset pricing factor. Even this requirement of connecting with marginal utility was removed with the invention of production-based asset pricing (Berk et al. (1999)), which models discount factors in reduced form, and allows most firm characteristics to be connected to risk and return. At the same time, our data-mined predictors are, in a sense, restricted. They all come from public accounting variables, which are, by construction, data points investors want to know. They also are restricted to simple functions of two variables, in contrast to the functions of many variables often seen in finance journals.

As a result, whether peer-reviewed theory places a meaningful control on data-mining bias is an empirical question—a question not addressed in Cochrane (2005); Harvey et al. (2016); or Harvey (2017). Our tests bring this question to the data and find that the answer is, disappointingly, no.

For published predictability more broadly, we find that peer-reviewed returns behave quite similarly to data-mined returns. Both published and data-mined returns increase in the years just before the original samples end, fall significantly in the first five years out-of-sample, and flatten out for years 10-15, before dipping temporarily around the 18th year out-of-sample. These patterns are not extracted from the matching process—the match is formed on only two in-sample summary statistics. Instead, these patterns emerge from the data itself: historical waves of finance publications and return predictability jointly produce the same patterns of portfolio returns, whether the portfolios come from peer review or data mining. It's as if the finance academics are just mining accounting data for return predictability,

[1]On using risk-based theory to discipline asset pricing tests, Cochrane (2009) writes: "these are the only standards we have to guard against fishing. In my opinion, the best hope for finding pricing factors that are robust out of sample and across different markets, is to try to understand the fundamental macroeconomic sources of risk."

and then decorating the results with stories about risk and psychology.

Replication code and the returns of 30,000 data-mined strategies can be found via `https://github.com/chenandrewy/flex-mining`. The predictor categorizations, as well as the excerpts that lead to the categorizations, are found at `https://github.com/chenandrewy/flex-mining/blob/main/DataInput/SignalsTheoryChecked.csv`.

1.1 Related Literature

Measuring peer-reviewed text provides a new angle on the long-standing debate on risk vs mispricing in the cross-section of stock returns (Fama (1970); Shiller (2003); Cochrane (2017); Barberis (2018); etc). Since Fama (1970), it has been recognized that standard empirical tests can only reject a specific special case of the broad class of risk theories. Our methods attack this problem by building on the efforts of the asset pricing community. This community is effectively an organic computer designed to explore the entire class of theories. Combined with our out-of-sample tests, our study of peer-reviewed texts shows that this massive computer does not find robust risk-based predictability.

We add to the literature on data mining in cross-sectional return predictability. Going back to Jensen and Benington (1970), researchers have been worried that data mining could result in finding lucky patterns that vanish out-of-sample. Lo and MacKinlay (1990) provides an early theoretical examination; Sullivan, Timmermann, and White (1999, 2001) study bias in market timing strategies; and McLean and Pontiff (2016) measure data-mining bias in published predictors. But to our knowledge, it was not until Yan and Zheng (2017) that anyone systematically mined accounting data for cross-sectional predictors.

Surprisingly, Yan and Zheng find that find data mining leads to "many" predictors that cannot be accounted for by luck using a bootstrap procedure (Fama and French (2010)). Moreover, they find data mining generates substantial out-of-sample alphas. We replicate and extend Yan and Zheng's results. While Yan and Zheng's data-mining strategies are inspired by functional forms used in the literature and are rescaled using economic intuition, our data-mining process is arguably free of economics. We show that not only does economics-free data mining generate substantial out-of-sample performance, but this out-of-sample performance is just as strong as the performance found through the peer-review process.

These results provide clarity to the conflicting evidence on accounting-based data mining. Using FDR methods, Harvey and Liu (2020) find evidence inconsistent with Yan and Zheng's results, while Chen (2022) finds evidence in support. Since FDR methods are complex and can be easily misinterpreted (Chen

and Zimmermann (2022b)), we focus exclusively on out-of-sample tests, which are well-understood and have straightforward interpretations. Our findings support not only Yan and Zheng's conclusion of "many" true predictors, but that the number of true predictors is in the thousands. In contemporaneous work, Goto and Yamada (2022) also find support for this conclusion.

Our results provide an alternative perspective on the question of how investors interact with academic research. McLean and Pontiff (2016) argue that investors learn from academic publications, as evidenced by the systematic decline in return predictability after publication that cannot be explained by data mining bias (see also Chen and Zimmermann (2020); Jensen et al. (2022)). This story is also seen in the trades of short sellers and hedge funds (McLean and Pontiff (2016); Calluzzo et al. (2019); McLean et al. (2020)). Our findings suggest that both academic research and investors are responding to the same fundamentals: the appearance of statistically significant return predictability in accounting data. Once return predictability appears, it is diminished through investor learning, and academics scientifically document a select subset of these phenomena.

2 Peer-Reviewed Theory and Out-of-Sample Performance

This section describes how we measure peer-reviewed theory. We also show how out-of-sample predictability varies by type of theory. Readers eager to compare theory with data mining should skip to Section 3.

2.1 Published Predictor Data

Our published cross-sectional predictors come from the Chen and Zimmermann (2022a) (CZ) dataset. This dataset is built from 207 firm-level variables that were shown to predict returns cross-sectionally in finance, accounting, and economics journals.

These variables cover the vast majority of firm-level predictors that can be created from widely-available data and were published before 2016. It covers 97, 90, 88, and 100 percent of predictors that were clearly shown to attain long-short significance using common datasets and are mentioned in McLean and Pontiff (2016); Harvey et al. (2016); Green et al. (2017); and Hou et al. (2020); respectively. These meta-studies, in turn, aim to provide comprehensive coverage of cross-sectional return studies.

We drop five predictors that produce mean long-short returns of less than 15 bps per month in-sample in CZ's replications. These predictors are rather distant from the original papers, and dropping

them ensures that the decay we document accurately reflects the literature.[2] Since these predictors are rare, including them has little effect on our results.

We drop another 10 predictors that have less than 9 years of post-sample returns. Most of these predictors rely on specialized data that have been discontinued, though a few are published relatively recently. This filter makes the out-of-sample results easy to interpret. But since the median post-sample length is about 20 years, including these predictors has little effect on our results.

Of the 191 predictors we examine, 67% were published in the Journal of Finance, Review of Financial Studies, Journal of Financial Economics, Journal of Financial and Quantitative Analysis, Review of Finance, or Management Science. 22% were published in top accounting journals (the Accounting Review, the Journal of Accounting and Economics, the Review of Accounting Studies, and the Journal of Accounting Research). The remaining 11% were published in a wide variety of economics, finance, and accounting journals, including the Journal of Political Economy, Quarterly Journal of Economics, and Review of Economic Dynamics.

For measuring out-of-sample performance, we use the "original paper" version of the CZ data. These data consist of long-short portfolios constructed following the procedures in the original papers. This choice is important, as out-of-sample decay varies by the details of the trading strategy (Chen and Velikov (2022)). Choosing the original implementations means that the decay we find is not due to a dispute with the peer review process about where exactly risk premiums should show up.

A caveat about the CZ data is that it includes only papers that were shown to predict firm-level returns using either portfolio sorts or regressions of returns on lagged characteristics. This omits many papers that only run contemporaneous regressions or GMM tests. Many such tests are designed to "price test assets," and so are common in the risk-based literature (e.g. Chen, Roll, and Ross (1986)). These omitted papers comprise roughly 1/3 of the 433 variables in Harvey, Liu, and Zhu's (2016) spreadsheet.[3]

---

[2]For example, CZ equal-weight the Frazzini and Pedersen (2014) betting against beta portfolios instead weighting by betas. CZ use CRSP age rather than the NYSE archive data used by Barry and Brown (1984). CZ also find very small returns in simple long-short strategies for select variables shown by Haugen and Baker (1996), Abarbanell and Bushee (1998), Soliman (2008) to predict returns in multivariate settings.

[3]A detailed list of these omitted variables can be found at `https://github.com/OpenSourceAP/CrossSection/blob/master/Comparison_to_HLZ.csv`.

2.2 Measuring Peer-Reviewed Theory

To classify predictors, we read the corresponding paper and identify a passage of text that summarizes the main argument. These passages are typically taken from either the abstract, introduction, or conclusion. We then categorize each argument as "risk," "mispricing," or "agnostic." Each predictor was reviewed by two of the authors to prevent errors.

Table 1 provides representative passages for predictors in each category. Risk-based passages are straightforward. These passages typically discuss risk or equilibrium, though a few also emphasize market efficiency. Mispricing passages discuss mispricing or investor errors. Agnostic passages either provide arguments for both sides or avoid discussing either issue.

Our analysis finds a remarkable consensus about the origins of cross-sectional predictability. This consensus is seen in Table 2, which counts the number of predictors in each theory category. Only 18% of cross-sectional predictors are judged by the peer review process to be due to risk. In contrast, 59% of predictors are due to mispricing. The remaining 23% of predictors are agnostic.

As a check on our manual classifications, we use software to count the ratio of "risk words" to "mispricing words" in each paper. For example, we count "utility," "maximize," and "priced" as risk words, and "behavioral," "optimistic," and "sentiment" as mispricing words (see Appendix A for a full list). Table 2 shows order statistics of this ratio within each manually-classified theory category. The median ratio for risk-based predictors is 4.00—that is, risk words appear four times more frequently than mispricing words. Mirroring this result, mispricing-based predictors have a median ratio of 0.20, indicating five times as many mispricing words. Overall, this simple word count supports our manual categorizations. The distribution of risk to mispricing words for risk-based predictors is far to the right of the other categories.

The word counts also support our finding that risk explains a small minority of predictors. Across all papers, the median risk-to-mispricing word ratio is 0.33, meaning that mispricing-related words are typically mentioned 3 times as frequently as risk-related words.

The consensus in Table 2 is perhaps surprising given the tone in recent reviews on empirical cross-sectional asset pricing (e.g. Bali et al. (2016); Zaffaroni and Zhou (2022)). These reviews provide a largely agnostic description of the origins of predictability, suggesting that peer review has come to a divided view. Our results show that the literature favors mispricing, and that only a small minority of predictors are due to risk, as judged by the community of finance scholars.

The sub-sample counts in Table 2 suggest an explanation for this recent agnosticism. Risk has been

Table 1: Peer-Reviewed Risk and Mispricing Examples

These examples illustrate how we manually categorize predictors as risk, mispricing, or agnostic. Risk-to-mispricing words are counted by software and defined in Appendix A.

Reference	Predictor	Example Text	Risk to Mispricing Words
Panel (a): Risk-Based			
Tuzel 2010	Real estate holdings	Firms with high real estate holdings are more vulnerable to bad productivity shocks and hence are riskier and have higher expected returns.	17.60
Bazdresch, Belo, and Lin 2014	Employment growth	We interpret this difference in average returns, which we refer to as the hiring return spread, as reflecting the relatively lower risk of the firms with higher hiring rates	7.32
Fama and MacBeth 1973	CAPM beta	The pricing of common stocks reflects the attempts of risk-averse investors to hold portfolios that are "efficient" in terms of expected value and dispersion of return.	2.31
Panel (b): Mispricing			
Ikenberry, Lakonishok, Vermaelen 1995	Share repurchases	Thus, at least with respect to value stocks, the market errs in its initial response and appears to ignore much of the information conveyed through repurchase announcements	0.05
Eberhart, Maxwell, and Siddique 2004	Unexpected R&D increase	We find consistent evidence of a misreaction, as manifested in the significantly positive abnormal stock returns that our sample firms' shareholders experience following these increases.	0.05
Desai, Rajgopal, Venkatachalam 2004	Operating Cash flows to price	CFO/P is a powerful and comprehensive measure that subsumes the mispricing attributed to all the other value-glamour proxies.	0.05
Panel (c): Agnostic			
Banz 1981	Size	To summarize, the size effect exists but it is not at all clear why it exists.	1.93
Boudoukh et al. 2007	Net Payout Yield	We show that the apparent demise of dividend yields as a predictor is due more to mismeasurement than alternative explanations such as spurious correlation, learning, etc.	1.00
Chordia, Subra, Anshuman 2001	Volume Variance	However, our findings do not lend themselves to an obvious explanation, so that further investigation of our results would appear to be a reasonable topic for future research.	0.21

Table 2: Risk or Mispricing? According to Peer Review

We categorize predictors into "risk," "mispricing," or "agnostic" based on manually reading the original papers (Table 1). We split the sample at 2004, around the advent of production-based asset pricing (Berk, Green, and Naik (1999); Gomes, Kogan, and Zhang (2003)). "Risk Words to Mispricing Words" shows the ratio of word counts in the papers. The word list is in Appendix A. p05, p50, and p95 are the 5th, 50th, and 95th percentiles within each theory category.

Source of Predictability	Num Published Predictors			Risk Words to Mispricing Words		
	Total	1981-2004	2005-2016	p05	p50	p95
Risk	35	4	31	0.33	4.00	12.76
Mispricing	113	49	64	0.07	0.22	1.21
Agnostic	43	15	28	0.12	0.49	3.81
Any	191	68	123	0.07	0.33	7.09

gaining popularity in the peer-review process recently, perhaps following the advent of production-based asset pricing (Berk, Green, and Naik (1999); Gomes, Kogan, and Zhang (2003)). While only 6% of the predictors published before 2005 were attributed to risk, this share increased quadrupled after 2005, to 25%.

This increase in risk-based predictors could be interpreted as scientific progress. Perhaps researchers had been searching in the wrong subspace of risk-based theories for decades after Treynor (1962). Indeed, the computational power required to solve Zhang's (2005) equilibrium model of the value premium may have been hard to find until the early 2000s.

However, science can also arrive at false paradigms (Kuhn (1962)). These false ideas can be especially difficult to remove if the evidence lacks the power to eliminate theories (Akerlof and Michaillat (2018)). This problem is relevant for risk-based asset pricing given the joint hypothesis problem, the generality of the SDF framework, and the observational nature of finance. We use large-scale out-of-sample tests to address these issues.

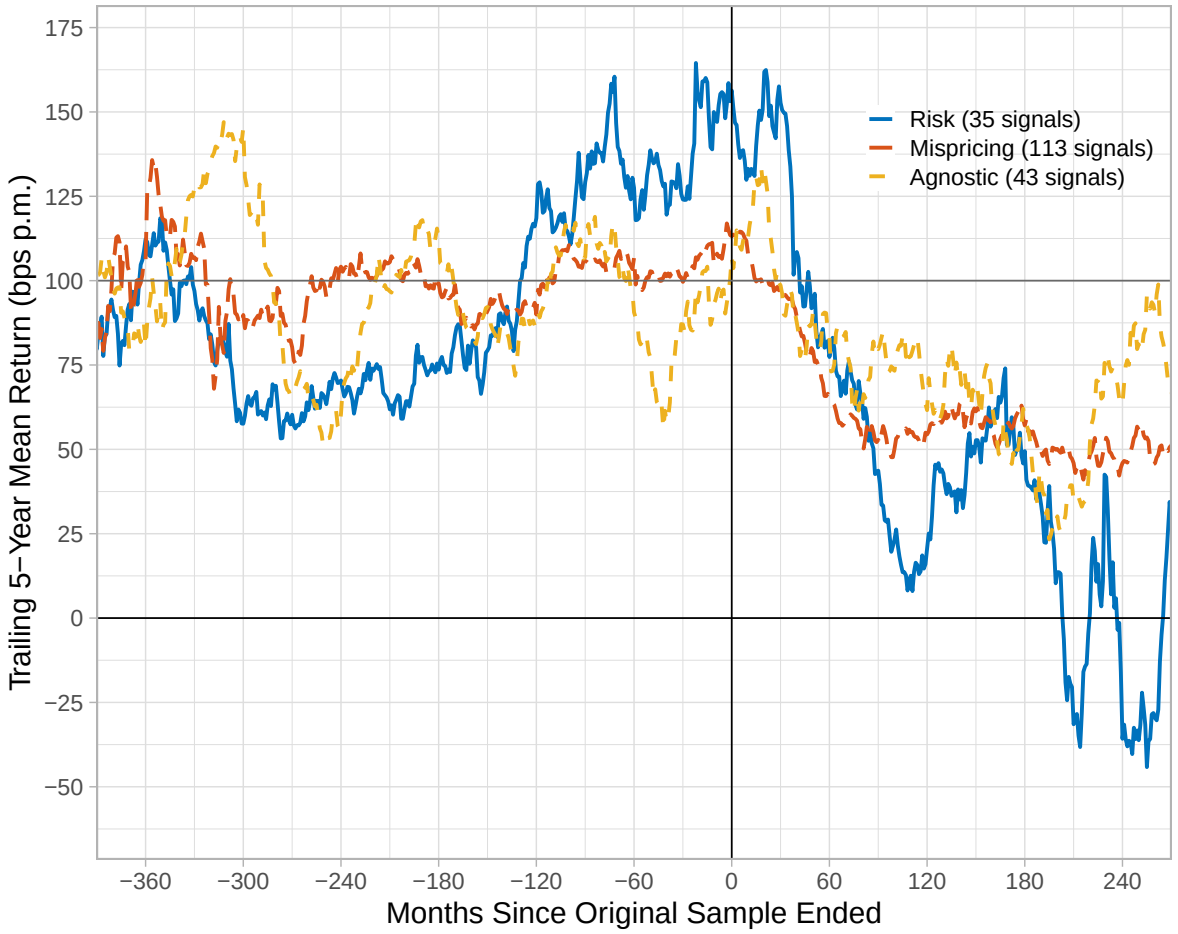
2.3 Out-of-Sample Performance by Peer-Reviewed Theory

In asset pricing, out-of-sample tests are perhaps the closest test we have to laboratory experiments. If a theory of predictability is true and stable, then predictability should be seen both in-sample and out-of-sample. Stability is a core theme of risk-based theory, which is based on the concept of economic equilibrium. Indeed, many of the risk-based predictors are based on infinite horizon equilibrium models. If these theories are approximately correct, then their predictions should hold for decades out of sample.

Figure 1 examines whether risk-based predictability holds out-of-sample. It plots the mean long-short returns of risk-based predictors (solid line) in event time, where the event is the end of the original papers' in-sample periods. We average across predictors within each month and then take the trailing 5-year average of these returns for ease of reading. Each strategy is normalized so that its mean in-sample return is 100 bps per month.

Figure 1: Out-of-Sample Returns by Peer-Reviewed Theory

The plot shows the mean long-short returns of published predictors in event time, where the event is the end of the original sample periods. Each predictor is normalized so that its mean in-sample return is 100 bps per month. Predictors are grouped by theory category based on the arguments in the original papers (Table 1). We average returns across predictors within each month and then take the trailing 5-year average for readability. Risk-based predictability decays notably, with returns centered around zero in months that are 7+ years out-of-sample.



Risk-based predictability decays quickly out-of-sample. In the first five years out-of-sample, the trailing five-year return plummets, from 150 bps per month to 75 bps. 10 years out-of-sample, the trailing mean return drops to 12 bps, and it continues to hover near zero for the rest of the out-of-sample

horizon.

Mispricing (dotted) and agnostic (dashed) predictors perform somewhat better. For these predictors, the trailing mean declines by about 25 percentage points in the first five years out-of-sample. Further out, the decay is roughly 50%.

Table 3 examines decay in a regression framework (following McLean and Pontiff (2016)). Specification (1) regresses monthly long-short returns on a post-sample indicator and its interaction with an indicator for risk-based theory. Returns are normalized to be 100 bps per month in-sample, so the post-sample coefficient implies that returns decay by 45 percent overall (across all theories). The interaction coefficient implies that risk-based theory leads to an additional decay of 13 percentage points, for a total decay of 58 percent.

The additional decay of risk predictors is far from statistically significant, with a standard error of 14 bps. Despite the minimum of 9 years of post-sample returns, the fact that peer-review only attributes 35 predictors to risk means that the data on this interaction is somewhat limited.

Nevertheless, there is plenty of data to show that risk-based predictors decay. This test is shown in the row "Null: Risk No Decay," which tests the hypothesis that the sum of the Post-Sample and Post-Sample $\times$ Risk coefficients is non-negative. The test rejects this hypothesis at the 1% level.

Specifications (2)-(4) show robustness. Specification (2) adds a post-publication indicator, specification (3) adds an indicator for mispricing-based theory, and specification (4) adds both. All three alternative specifications arrive at risk-based predictors decaying by an additional 15 to 20 percentage points. Specification (4) implies that post-publication, being risk-based implies an additional 14 + 7 = 21 percentage points of decay, for a total decay of 22 + 22 + 21 = 65%. In the Appendix, we show that decay occurs even for predictors with the highest risk words to mispricing words ratio (Figure B.1).

One might expect mispricing-based predictability to decay post-publication, as investors learn about mispricing. After a journal publicizes a particular underpricing, investors may buy the underpriced stocks, and trade away the expected returns (McLean and Pontiff (2016)). The row "Null: Mispricing Decay" shows that one cannot reject this hypothesis, with a p-value that is essentially 1.0.

A natural question, then, is whether investors learn about risk. Just as investors may learn about undervalued stocks from journals, they may also learn that certain stocks expose them to macroeconomic risk. They might then avoid these stocks and demand a higher risk premium, which would show up in a higher post-publication return. The row "Null: Risk No Decay" shows that one can strongly reject this hypothesis, with a p-value less than 0.1%.

Table 3: Regression Estimates of Theory Effects on Predictability Decay

We regress monthly long-short strategy returns on indicator variables to quantify the effects of peer-reviewed theory on predictability decay. Each strategy is normalized to have 100 bps per month returns in the original sample. "Post-Sample" is 1 if the month occurs after the predictor's sample ends and is zero otherwise. "Post-Pub" is defined similarly. "Risk" is 1 if peer review argues for a risk-based explanation (Table 1) and 0 otherwise. "Mispricing" is defined similarly. Parentheses show standard errors clustered by month. "Null: Mispricing Decay" shows the p-value that tests whether mispricing-based returns decrease post-sample (Column (3)) or post-publication (Column (4)). "Null: Risk No Decay" shows the p-value that tests whether risk-based returns do not decrease post-sample ((1) and (3)) or post-publication ((2) and (4)). The decay in risk-based predictors is highly statistically significant, and inconsistent with the hypothesis that risk theory uncovers stable expected returns.

RHS Variables	(1)	(2)	(3)	(4)
Intercept	100	100	100	100
	(6.4)	(6.4)	(6.4)	(6.4)
Post-Sample	-45.0	-23.5	-39.6	-22.2
	(8.3)	(11.8)	(10.2)	(15.8)
Post-Pub		-26.9		-21.5
		(11.6)		(17.9)
Post-Sample x Risk	-12.5	-12.4	-17.9	-13.7
	(14.1)	(23.9)	(15.7)	(25.7)
Post-Pub x Risk		-1.5		-6.9
		(29.2)		(32.5)
Post-Sample x Mispricing			-7.5	-1.8
			(7.5)	(16)
Post-Pub x Mispricing				-7.5
				(18)
Null: Mispricing Decay			1.0	1.0
Null: Risk No Decay	$< 0.1\%$	$< 0.1\%$	$< 0.1\%$	$< 0.1\%$

3 Peer-Reviewed Theory vs Naive Data-Mining

We've shown that risk-based predictors decay notably out-of-sample, underperforming mispricing and agnostic predictors. But is using risk-based theory at least better than using no theory at all? This section answers this question by comparing published strategies to matched data-mined strategies.

3.1 Data-Mined Trading Strategies

We generate 29,315 signals as follows. Let X be one of 242 Compustat accounting variables + CRSP market equity and Y be one of the 65 variables that is observed and positive for $> 25\%$ of firms in 1963 with matched CRSP data. We form signals by combining all combinations of ratios (X/Y) and scaled first differences ($\Delta X/\text{lag}(Y)$). This procedure would lead to $242 \times 65 \times 2 = 31,460$ signals, but we drop

2,145 signals that are redundant in "unsigned" portfolio sorts.[4]

We lag each signal by six months, and then form long-short decile strategies by sorting stocks on the lagged signals in each June. Delisting returns and other data handling methods follow Chen and Zimmermann (2022a) to ensure that the published and data-mined strategies are comparable. For further details, please see the Github repo.

In our view, this process is the simplest reasonable data mining procedure. A reasonable data mining procedure should include both ratios and first differences. Scaling first differences by a lagged variable nests percentage changes, which likely should also be included in a reasonable data mining process. This data mining procedure includes little, if any, economic insight.

This procedure is inspired by Yan and Zheng (2017), who create 18,000 signals by applying 76 transformations to 240 accounting variables. These transformations are inspired, in part, by the asset pricing literature. Choosing transformations based on the literature could, potentially, lead to look-ahead bias. Our procedure avoids this potential bias, though we find that the results change very little using Yan and Zheng's methods.

3.2 "Out-of-Sample" Performance of Data Mining

Table 4 shows that this simple procedure leads to substantial "out-of-sample" predictability. Starting in 1994, we sort strategies into five bins based on their past 30 years of return (in-sample). We then examine the return over the next year in each bin (out-of-sample). The table shows the average statistics for each bin, averaged across each year. We put "out-of-sample" in quotes here because this concept differs from the out-of-sample concept used in the rest of the paper.

The equal-weighted bin 1 returns -49 bps per month out-of-sample. The negative return is predictable: bin 1 is composed of the 5,800 strategies with the most negative in-sample returns. Indeed, bin 1's mean return in-sample is on average -59 bps per month, implying a mild decay of only 17% out-of-sample. A similar persistence is seen in bin 5, which decays by 27%. Bins 2 and 3 also show persistence, though the decay is larger. Bin 4 has, on average, returns very close to zero in-sample, so the percentage decay is not well defined, but its out-of-sample returns are also close to zero. Return persistence is also seen in value-weighted strategies, though the magnitudes are generally weaker. Still, the decay is far from

[4]For the $65 \times 65 = 4,225$ ratios where the numerator is also a valid denominator, there are only 65 choose 2 = 2,080 ratios that are distinct in the sense that there are no ratios which would lead to identical rankings if the sign was flipped.

Table 4: "Out-of-Sample" Returns of Data-mined Accounting Strategies

We summarize our 29,000 data-mined strategies using out-of-sample sorts. For each June starting in 1994, we sort strategies into 5 bins based on their past 30-year mean returns ("in-sample"). We then compute the mean return over the next year within each bin ("out-of-sample"). Statistics are calculated at the strategy level, then averaged within the bin, then averaged across sorting years. Decay is the percentage change toward zero in the mean return out-of-sample relative to in-sample. We omit decay for bin 4 because the mean return in-sample is negligible. Mean returns are bps per month. Naive data mining can generate performance comparable to the peer review process, both in- and out-of-sample. Data-mining predictability is weaker post-2003, especially in large stocks.

Panel (a): Full Sample

In-Sample	Equal-Weighted Long-Short Deciles				Value-Weighted Long-Short Deciles			
	Past 30 Years (IS)		Next Year (OOS)		Past 30 Years (IS)		Next Year (OOS)	
Bin	Mean Return	t-stat	Mean Return	Decay (%)	Mean Return	t-stat	Mean Return	Decay (%)
1	-59.3	-4.24	-49.4	16.7	-37.6	-2.06	-16.3	56.6
2	-29.1	-2.46	-18.9	35.1	-15.7	-1.02	-5.6	64.0
3	-13.3	-1.20	-3.2	75.9	-4.9	-0.33	-1.8	62.7
4	-0.3	-0.04	5.6		5.4	0.35	-0.0	
5	23.4	1.46	17.1	26.9	27.1	1.37	10.8	60.3

Panel (b): Bins Sorted 2003-2019

In-Sample	Equal-Weighted Long-Short Deciles				Value-Weighted Long-Short Deciles			
	Past 30 Years (IS)		Next Year (OOS)		Past 30 Years (IS)		Next Year (OOS)	
Bin	Mean Return	t-stat	Mean Return	Decay (%)	Mean Return	t-stat	Mean Return	Decay (%)
1	-59.9	-4.04	-27.1	54.8	-37.2	-1.88	-4.6	87.6
2	-28.4	-2.32	-10.6	62.7	-14.4	-0.90	-0.7	95.1
3	-11.6	-1.00	-0.1	99.1	-3.8	-0.25	-1.7	55.5
4	2.3	0.19	7.8		6.1	0.40	-3.3	
5	25.9	1.53	17.4	32.8	28.4	1.38	2.3	91.9

zero and in the ballpark of the out-of-sample decay for published strategies (McLean and Pontiff (2016)). These results extend the findings of Yan and Zheng (2017), who show that simple data mining can generate out-of-sample alpha.

Return predictability is also seen in the in-sample t-statistics, which are quite far from the null of no predictability. If the t-stats were standard normal (as implied by the central limit theorem), then the mean t-stat in bin 1 would be -1.40.[5] In contrast, the equal-weighted bin 1's mean in-sample t-stat is -4.2.

[5]The t-stats in bin 1 would be standard normal truncated above at $\Phi^{-1}(0.2) = -0.84$, where $\Phi^{-1}(\cdot)$ is the inverse standard

This bin contains 5,800 trading strategies, implying a very large number of predictors that exhibit true predictability. In value-weighted strategies, bin 1 has a mean t-stat of -2.1, moderately larger than the null of 1.40. As shown in Yan and Zheng's (2017), bootstrap test, the difference between the in-sample t-stats and the null cannot be accounted for by luck.

Panel (b) of Table 4 describes data-mined strategies in the more recent sub-sample. Data-mined return predictability has declined significantly post-2003, consistent with the decline in published return predictability found by Chordia et al. (2014); McLean and Pontiff (2016); and Chen and Velikov (2022). The decay of the strongest equal-weighted strategies is roughly 50% post-2003, compared to about 20% for the full sample. Post-2003, value-weighted strategy decay is not far from 100%, consistent with Chen and Velikov's (2022) finding that anomaly returns are essentially zero after trading costs in recent years. These results emphasize that decay is not just due to the publicization of anomalies—the improvements in liquidity and information technology around the early 2000s also seem to play a critical role. These results also mean that it is important to carefully match sample periods when comparing published and data-mined strategies.

3.3 Matching Data-Mined Predictors to Peer-Reviewed Predictors

Our matching addresses the following question: Suppose you have a predictor with a given in-sample mean return and t-stat. How should your views on out-of-sample returns change if you learn that the predictor is data-mined instead of based on peer-reviewed theory?

The matching proceeds as follows: For each published predictor, we find all data-mined predictors with the same stock weighting (equal- or value-weighted), absolute t-stats within 10 percent, and absolute mean returns within 30 percent, all calculated using the published predictor's in-sample period. We also require that the data-mined strategy has 12 observations in the last year of the in-sample period and at least 20 stocks every month during the full in-sample period (excluding months before the signal was available). We then average across all matched strategies to form a data-mined benchmark for each peer-reviewed predictor.

Table 5 describes the match. The top panel shows that matched predictors are quite close to peer-reviewed predictors in terms of mean in-sample statistics. For each theory category, the mean matched

normal CDF. The mean of this truncated normal is $-\phi(-0.84)/[\Phi(-0.84)] = -1.40$. This calculation also assumes weak dependence across t-stats.

Table 5: Summary of Matching Peer-Reviewed to Data-Mined Predictors

For each peer-reviewed predictor, find data-mined predictors that have absolute t-stats 10% and absolute mean returns within 30%, using the peer-reviewed sample periods. The top panel shows mean returns and t-stats, averaged within peer-reviewed theory categories. For the matched predictors, we average within each peer-reviewed predictor and then average across peer-reviewed predictors. The bottom panel shows the number of matches for each peer-reviewed predictor. Naive data-mining readily generates in-sample mean returns and t-stats comparable to those that come from peer review. Most peer-reviewed predictors have more than 100 data-mined counterparts.

Source of	Median In-Sample Period		Mean Return (IS)		t-stat (IS)	
Predictability	Start	End	Published	Matched	Published	Matched
Risk	1965	2003	55.4	50.0	3.51	3.45
Mispricing	1975	1999	69.8	63.1	3.85	3.79
Agnostic	1965	2002	61.9	54.4	3.47	3.45

Source of	Number of matched strategies per predictor					Unmatched	Matched
Predictability	Min	25th	50th	75th	Max	Predictors	Predictors
Risk	19	260	564	727	990	2	33
Mispricing	1	152	362	683	1140	8	105
Agnostic	42	317	569	819	1203	2	41

data-mined t-stat is within 0.06 of the published strategies, and the mean in-sample return is within 8 bps.

The bottom panel shows that finding matches is quite easy. Most peer-reviewed predictors have more than 100 matches in the data-mined data. The median peer-reviewed predictor has several hundred matches, suggesting that data mining isn't simply recovering the peer-reviewed predictor. This result shows that theory is not necessary for finding strong in-sample performance. We will soon see that it is also not necessary to find strong out-of-sample performance.

Out of the 191 published predictors, 12 remain unmatched. Most of the unmatched predictors obtain extremely high t-stats using non-accounting data. For example, Yan's (2011) put volatility minus call volatility predictor uses option prices and Hartzmark and Soloman's (2013) dividend seasonality uses CRSP dividend payments to achieve t-stats of 7.7 and 14.4, respectively. These results imply that adding more datasets to the data mining process would lead to a near-complete matching, though the benefit may not be worth the cost, given the relatively small number of unmatched predictors. The Appendix provides the complete list (Table B.1).

3.4 Out-of-Sample Returns of Peer-Reviewed vs Data-Mined Predictors

Figure 2 compares the out-of-sample performance of peer review and data mining. It plots the mean returns of each class of predictor in event time, where the event is the end of the published predictors' in-sample periods. Data-mined strategies are signed to have positive in-sample returns and all strategies are normalized to have 100 bps return in-sample. The figure averages across predictors within each event-time month and then takes the trailing 5-year average to smooth out noise.

Peer review and data mining have eerily similar event time returns. The data-mined returns (dot-dash) resemble a Kalman-filtered version of the peer-reviewed returns (solid). The peaks and troughs broadly match for both series, throughout the event time horizon. This detailed fit is *not* coming from the matching process. We match only on the mean returns and t-stats through the whole in-sample period, ignoring any patterns within the sample periods. This commonality is a property of the accounting and returns data itself, and the way the data interacts with peer-reviewed research.

Out-of-sample, peer-reviewed and data-mined predictors perform similarly. For both groups, the trailing 5-year return increases to about 120 bps per month just as the sample ends, and then drops to around 60 bps per month five years after the sample ends. For both groups, returns hover around 40-60 bps per month for the remainder of the event time horizon.

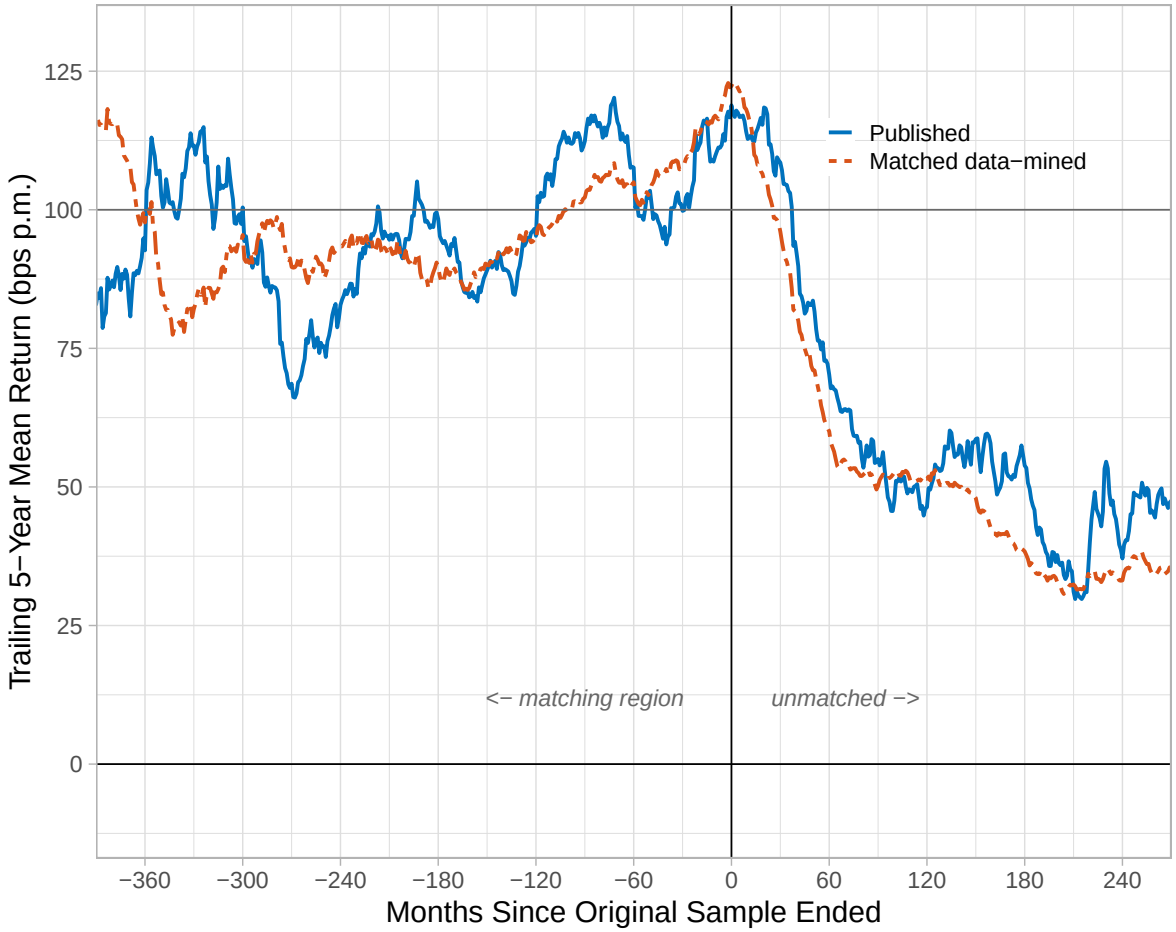
These results imply that data mining works just as well as reading peer-reviewed journals. Back-testing accounting signals, unguided by theory, leads to the same out-of-sample returns as drawing on the best ideas from the best finance departments in the world. We emphasize that these accounting signals are very simple functions and are selected with the simplest of statistical methods. A typical finance undergraduate should be able to understand these methods, though it may take a bit of computer science training to code up the algorithm.

Of course, academic publications can destroy return predictability by publicizing mispricing (McLean and Pontiff (2016)). So the similar performance in Figure 2 may be due to offsetting effects. It could be that peer-reviewed predictability would have out-performed, if not for the publicization of mispricing.

Figure 3 zooms in these results by separating out predictors by peer-reviewed theories. Panel (a) shows only predictors that are, according to peer review, due to risk. These predictors should not have offsetting effects related to the elimination of mispricing—if the peer-reviewed theories are correct. However, Panel (a) shows predictors founded in equilibrium theory perform no better than data mining. Risk-based predictability appears stronger in the first few years out-of-sample, but this outperformance vanishes around year 7. Since the publication dates are typically 4 years after the original samples end,

Figure 2: Out-of-Sample Returns of Peer-Review Predictors vs a Data-Mined Benchmark

The plot shows the mean long-short returns of published predictors in event time, where the event is the end of the original sample periods. All predictors are normalized to have 100 bps mean return in-sample. Returns are averaged across predictors within each month, and then the trailing 5-year average is taken for readability. The solid line shows published predictors, dotted shows matched data-mined predictors. Data-mined predictors come from building ratios or scaled first differences of 240 accounting variables as described in Section 3.1. Matching is described in Table 5. Naive data mining leads to out-of-sample returns that are eerily similar to the peer-review process.



it is not until around year 7 that the trailing 5-year mean return is fully out-of-sample.

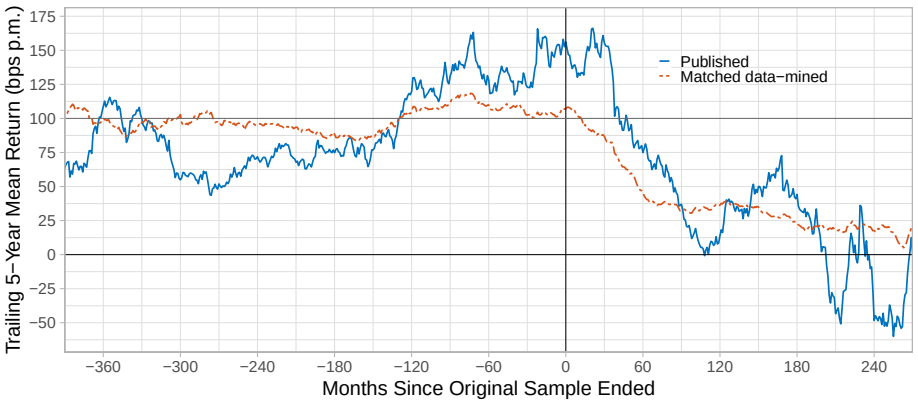
This result may be surprising, given the textbook idea that rigorous economic theory is our best protection against data mining bias (Cochrane (2009); Harvey et al. (2016)). This idea follows from a simple logic: data mining bias arrives from selecting the best results from a large set (Chen and Zimmermann (2022b)). Economic theory constrains this set, limiting the selection bias.

However, the set of predictors allowed by peer-reviewed theory is quite broad. Since at least Fama (1970), finance academics have recognized that there are many possible theories of risk, and thus many possible risk-based predictors. Even if one specifies a theory of risk, the key objects are often not

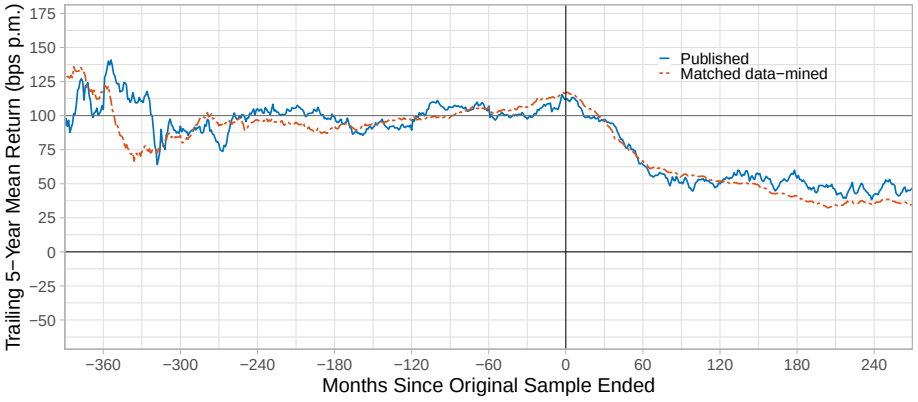
Figure 3: Peer-Review vs Data-Mining by Theoretical Justification

The plot shows the mean long-short returns of published predictors in event time, where the event is the end of the original sample periods. Predictors are normalized to have 100 bps mean return in-sample. Data-mined predictors come from building ratios or scaled first differences of 240 accounting variables as described in Section 3.1. Matching is described in Table 5. Risk-based theory does no better than data-mining for generating out-of-sample returns.

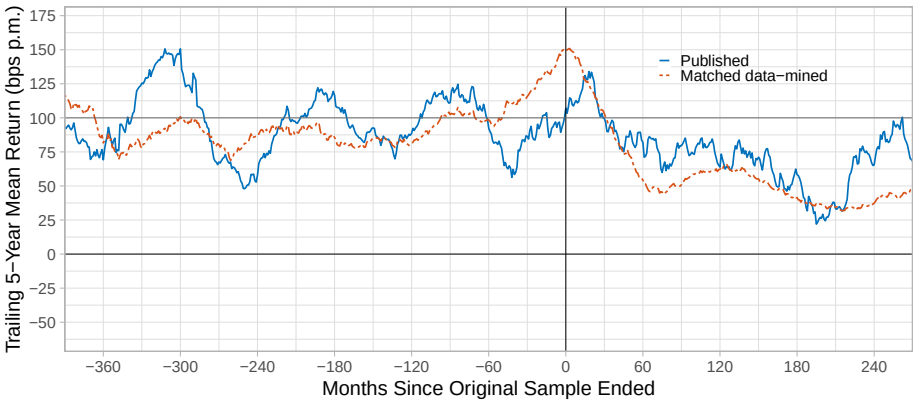
(a) Risk-Based



(b) Mispricing-Based



(c) Agnostic



observed (Roll (1977)). These objects must be proxied using various empirical measures, with each empirical measure potentially leading to a different predictor. The Merton ICAPM and production-based asset pricing greatly expand both the sets of mathematical theories that are examined as well as the possible proxies. Moreover, the calibration philosophy that asset pricing inherited from macroeconomics means that peer review can be quite generous when judging the empirical validity of a theory. Thus economic theory, as it is practiced in journals, may place little restriction on the set of predictors being tested. And given the popularity of equilibrium theory in Ph.D. coursework and Nobel prizes, it is likely that this set is pushed as far as possible by researchers seeking tenure and citations.

At the same time, the set of data-mined predictors could be considered restricted. The data are all composed of public accounting variables, which are by construction variables that investors want to know for valuing stocks. Moreover, the space of functions used in our data mining procedure is quite small. We use only two functions of two accounting variables, with no parameters. It should perhaps not be surprising, then, that our data-mining exercise is similar to peer-reviewed theory in terms of out-of-sample performance.

Panels (b) and (c) of Figure 3 examine mispricing and agnostic predictors, respectively. For both types of predictors, out-of-sample predictability is similar to that obtained from naive data mining. For the mispricing-based predictors, the published (solid) and data-mined (dot-dash) lines are so similar it looks as if they are all operating on the same underlying mechanism. The agnostic predictors outperform, but the difference is comparable to the standard error of roughly 20 bps per month seen in Table 3.

Overall, the similarity between data-mined and published returns suggests a different view of the McLean and Pontiff (2016) facts. McLean and Pontiff argue that investors learn about mispricing from academic publications, as seen in the fact that predictability systematically decays after publication. But investors are surely learning from the accounting data itself. One wonders, then, how much academics contribute. Figure 2 suggests that the contribution is minor. It looks as if both academics and investors are learning from the accounting data in parallel. Once evidence of predictability becomes strong enough, both investors and academics act, the former to correct the mispricing, and the latter to document it scientifically.

3.5 A Closer Look at Value and Momentum

How, exactly, does data mining achieve journal-like out-of-sample returns? Tables 6 and 7 take a closer look, by listing the data-mined predictors that matched with Fama and French's (1992) B/M and Jegadeesh and Titman's (1993) 12-month momentum.

Table 6: 20 Data-Mined Predictors With Returns Similar to Fama-French's B/M (1992)

Table lists 20 of the 171 data-mined signals that performed similarly to Fama and French's (1992) B/M in the original 1963-1990 sample period in terms of mean returns and t-stats. Signals are ranked according to the absolute difference in mean in-sample return. Sign = -1 indicates that a high signal implies a lower mean return in-sample. Data mining picks up themes found by peer-reviewed research (e.g. investment, equity issuance) and leads to similar out-of-sample performance as Fama and French's B/M.

Similarity Rank	Signal	Sign	Mean Return (% Monthly) 1963-1990	1991-2021
Peer-Reviewed				
	Book / Market (Fama-French 1992)	1	0.96	0.57
Data-Mined				
1	Δ[Assets]/lag[Operating expenses]	-1	0.96	0.90
2	Δ[PPE net]/lag[Sales]	-1	0.96	0.80
3	[Market equity FYE]/[Depreciation & amort]	-1	0.95	0.66
4	Δ[Assets]/lag[Cost of goods sold]	-1	0.95	0.85
5	Δ[Assets]/lag[SG&A]	-1	0.95	0.82
6	Δ[PPE net]/lag[Gross profit]	-1	0.98	0.51
7	Δ[PPE net]/lag[Current liabilities]	-1	0.94	0.90
8	[Depreciation (CF acct)]/[Capex PPE sch V]	1	0.97	0.78
9	[Market equity FYE]/[Depreciation depl amort]	-1	0.94	1.03
10	[Stock issuance]/[Capex PPE sch V]	-1	0.94	0.93
	. . .			
101	[Market equity FYE]/[Current assets]	-1	1.15	0.89
102	[Market equity FYE]/[Common equity]	-1	1.14	0.51
103	Δ[Cost of goods sold]/lag[Cost of goods sold]	-1	0.75	0.94
104	Δ[Receivables]/lag[Invested capital]	-1	0.76	0.46
105	Δ[Receivables]/lag[PPE net]	-1	0.76	0.46
	. . .			
167	Δ[Assets]/lag[IB adjusted for common s]	-1	0.67	-0.02
168	Δ[Cost of goods sold]/lag[Current liabilities]	-1	0.67	0.68
169	Δ[Long-term debt]/lag[Operating expenses]	-1	0.67	0.50
170	Δ[Depreciation & amort]/lag[Invested capital]	-1	0.67	0.67
171	Δ[Current liabilities]/lag[Inventories]	-1	0.67	0.49
	Mean Data-Mined		0.83	0.69

Table 6 begins with B/M. At the top of the table, we see that predictors related to asset growth had in-sample performance extremely similar to B/M, as did predictors related to depreciation and equity issuance. Moving down the table, we see predictors that are somewhat more distant, but that still

achieved mean returns within 20 bps of B/M. These predictors include those related to cost growth and working capital investment. Still other predictors that performed similarly to B/M in-sample include one related to debt issuance.

Table 7: 20 Data-Mined Predictors That Perform Similarly to Jegadeesh and Titman's Momentum (1993)

Table lists 20 of the 44 data-mined signals that performed similarly to Jegadeesh and Titman's (1993) 12-month momentum in the original 1964-1989 sample period in terms of mean returns and t-stats. Signals are ranked according to the absolute difference in mean in-sample return. Sign = -1 indicates that a high signal implies a lower mean return in-sample. Data mining picks up themes found by peer-reviewed research (e.g. profitabitability, investment) and leads to similar out-of-sample performance as Jegedeesh and Titman's momentum.

Similarity Rank	Signal	Sign	Mean Return (% Monthly) 1964-1989	1990-2021
Peer-Reviewed				
	12-Month Momentum (Jegadeesh-Titman 1993)	1	1.38	0.49
Data-Mined				
1	[Retained earnings unadj]/[Market equity FYE]	1	1.38	-0.19
2	[Retained earnings unadj]/[Assets other sundry]	1	1.40	0.15
3	[PPE and machinery]/[Current liabilities]	1	1.42	0.38
4	[Retained earnings unadj]/[Cash & ST investments]	1	1.42	0.17
5	[PPE and machinery]/[Capital expenditure]	1	1.50	0.79
6	[Retained earnings unadj]/[Invest & advances other]	1	1.51	0.04
7	[Investing activities oth]/[Nonoperating income]	1	1.52	0.11
8	[Income taxes paid]/[PPE net]	1	1.22	0.14
9	[PPE and machinery]/[Capex PPE sch V]	1	1.56	0.80
10	[Market equity FYE]/[Current assets]	-1	1.20	0.88
	. . .			
21	[Depreciation (CF acct)]/[Market equity FYE]	1	1.10	0.73
22	[Retained earnings unadj]/[Common equity]	1	1.66	-0.08
23	Δ[Assets]/lag[Current assets]	-1	1.09	1.26
24	Δ[PPE (gross)]/lag[Operating expenses]	-1	1.09	0.70
25	[Funds from operations]/[Market equity FYE]	1	1.07	-3.51
	. . .			
40	[Deprec end bal (Sch VI)]/[Market equity FYE]	1	1.02	1.03
41	[Market equity FYE]/[Cost of goods sold]	-1	1.00	0.70
42	[Rental expense]/[Market equity FYE]	1	1.00	0.84
43	Δ[Invested capital]/lag[Current assets]	-1	0.97	1.32
44	[Retained earnings unadj]/[Receiv current other]	1	1.79	0.19
	Mean Data-Mined		1.29	0.48

Table 7 lists data-mined predictors that performed similarly to Jegadeesh and Titman's (1993) 12-month momentum. Many themes seen in Table 6 show up again in Table 7, though we also see profitability-related predictors, as well as some unusual variables (rental expense). The Appendix lists predictors related to Banz's (1981) Size predictor (Table B.2), which also include well-known themes

(investment and profitability) and more unusual variables (investment tax credits and interest expense).

Overall, the themes seen in Tables 6 and 7 echo those found in the cross-sectional predictability literature. One may have thought that linking investment or profitability to expected returns requires some economic insight, but it turns out that sheer data mining can systematically uncover these patterns. And while one may have thought that economic insight is required to find the out-of-sample robustness found in Fama and French's (1992) B/M, it turns out this is not the case. On average, the data-mined predictors in Table 6 returned 67 bps in the 30 years after Fama and French's sample, slightly higher than the 57 bps of B/M. Similarly, the data-mined counterparts to momentum earned 48 bps per month out-of-sample, very close to the 49 bps earned by momentum. Data-mined counterparts to Banz's (1981) Size also performed similarly (Table B.2).

4 Conclusion

We take stock of 45 years of cross-sectional asset pricing research by combining text analysis with out-of-sample tests. Our text analysis finds that only 20% of published cross-sectional stock return predictors are due to risk. Post-publication, the returns of the few risk-based predictors largely disappear. Risk-based predictors fail to outperform naively data-mined strategies out-of-sample.

These findings have strong negative implications for either risk-based theory or the peer-review process. If risk-based theory is true, then the peer-review process uncovers only false theories, or the subset of theories that largely vanishes out-of-sample. But if peer review is working well, then the entire class of risk-based theory is not helpful for understanding the cross-section of expected stock returns.

Either way, our finding that peer-reviewed results are no better than data mining at predicting out-of-sample returns has important implications for our understanding of asset prices. These results imply that cross-sectional asset pricing research, risk-based or otherwise, provides little incremental information in real-time. Though our findings are quite negative about theory, they suggest that data mining has been an underutilized tool for understanding financial markets.

Appendix A Risk words and mispricing words

We remove stopwords, lowercase and lemmatize all words using standard methods. Then, we count separately the words corresponding to risk and mispricing.

We consider as risk words the following terms and their grammatical variations: "utility," "maximize," "minimize," "optimize," "premium," "premia," "premiums," "consume," "marginal," "equilibrium," "sdf," "investment-based," and "theoretical." We also count as risk words appearances of “risk” that are not preceded by “lower,” and appearances of “aversion,” “rational,” and “risky” that are not preceded by “not.”

The mispricing words consist of " "anomaly," "behavioral," "optimistic," "pessimistic," "sentiment," "underreact," "overreact," "failure," "bias," "overvalue," "misvalue," "undervalue," "attention," "underperformance," "extrapolate," "underestimate," "misreaction," "inefficiency," "delay," "suboptimal," "mislead," "overoptimism," "arbitrage," "factor unlikely," and their grammatical variations. We further count as mispricing the terms "not rewarded," "little risk," "risk cannot [explain]," " low [type of] risk," "unrelated [to the type of] risk," "fail [to] reflect," and "market failure," where the terms in brackets are captured using regular expressions or correspond to stopwords.

Appendix B Additional Results

Figure B.1: Out-of-Sample Returns vs Risk to Mispricing Words

Each marker represents one published predictor's mean return. The regression line is fitted with OLS. The full reference for each acroynym can be found at `https://github.com/OpenSourceAP/CrossSection/blob/master/SignalDoc.csv`. Even predictors with the strongest focus on risk decay on average.

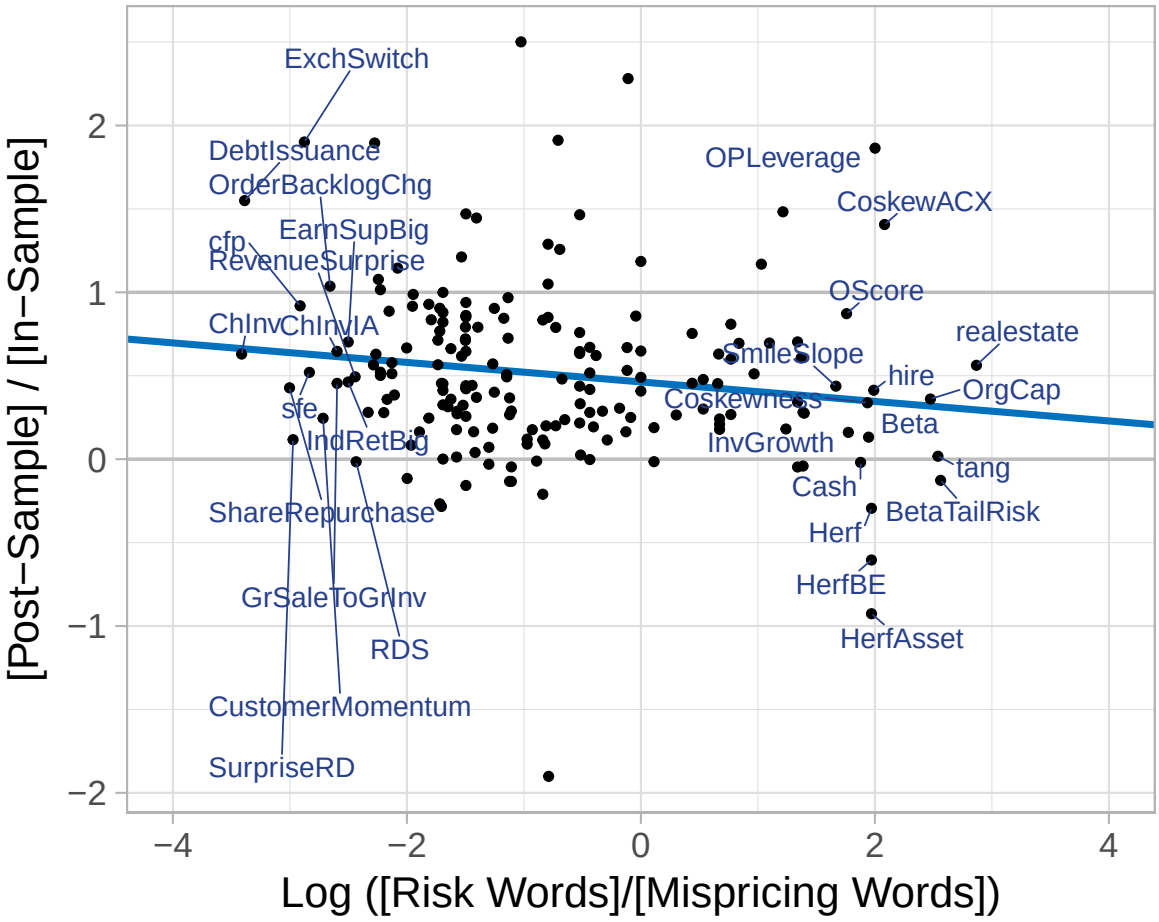


Table B.1: Unmatched Peer-Reviewed Predictors

We list peer-reviewed predictors that have zero matched data-mined accounting signals. All of the failed matches have extremely large in-sample t-stats. Most of the failed matches use non-accounting data (e.g. option prices, analyst forecasts), suggesting expanding the data-mined dataset would mostly complete the matching process, though the benefit may not be worth the cost.

Reference	Predictor	Theory	Mean Return	t-stat
Litzenberger and Ramaswamy (1979)	Predicted div yield next month	Risk	41.3	4.20
Yan (2011)	Put volatility minus call volatility	Risk	183.2	7.71
Asquith Pathak and Ritter (2005)	Inst own among high short interest	Mispricing	240.9	3.35
Chan, Jegadeesh and Lakonishok (1996)	Earnings announcement return	Mispricing	120.0	12.97
Hartzmark and Salomon (2013)	Dividend seasonality	Mispricing	32.8	14.39
Hou (2007)	Industry return of big firms	Mispricing	220.1	9.13
Loh and Warachka (2012)	Earnings surprise streak	Mispricing	109.3	10.52
Richardson et al. (2005)	Change in financial liabilities	Mispricing	72.7	12.10
Spiess and Affleck-Graves (1999)	Debt issuance	Mispricing	21.3	3.94
Zhang (2004)	Firm age - momentum	Mispricing	233.1	5.38
Jegadeesh (1989)	Short term reversal	Agnostic	292.5	14.20
Novy-Marx (2012)	Intermediate momentum	Agnostic	123.7	5.86

Table B.2: 20 Data-Mined Predictors That Perform Similarly to Banz's Size (1981)

Table lists 20 of the 221 data-mined signals that performed similarly to Banz's (1981) size in the original sample period. Signals are ranked according to the absolute difference in mean in-sample return. Sign = -1 indicates that a high signal implies a lower mean return in-sample. Data mining leads to similar out-of-sample performance.

Similarity Rank	Signal	Sign	Mean Return (% Monthly) 1926-1975	1976-2021
Peer-Reviewed				
	Size (Banz 1981)	-1	0.50	0.22
Data-Mined				
1	Δ[Equity liq value]/lag[Sales]	-1	0.50	0.77
2	Δ[Invest tax credit inc ac]/lag[Pref stock redemp val]	-1	0.49	-0.13
3	[Cost of goods sold]/[Capex PPE sch V]	1	0.50	0.82
4	Δ[Assets]/lag[Preferred stock liquidat]	-1	0.49	0.20
5	Δ[Equity liq value]/lag[Curr assets other sundry]	-1	0.48	0.73
6	Δ[Equity liq value]/lag[Current liabilities]	-1	0.48	0.86
7	Δ[Receivables]/lag[Pref stock redemp val]	-1	0.48	0.19
8	[Market equity FYE]/[Invested capital]	-1	0.49	0.73
9	Δ[Current assets]/lag[Invest tax credit inc ac]	-1	0.52	0.33
10	Δ[Assets]/lag[Pref stock redemp val]	-1	0.47	0.25
	. . .			
101	Δ[Sales]/lag[Current liabilities]	-1	0.40	0.77
102	Δ[Interest expense]/lag[Cost of goods sold]	-1	0.40	0.71
103	Δ[Interest expense]/lag[Num employees]	-1	0.40	0.68
104	Δ[Operating expenses]/lag[Long-term debt]	-1	0.40	0.48
105	Δ[Operating expenses]/lag[Current liabilities]	-1	0.40	0.90
	. . .			
234	[Gross profit]/[Earnings before interest]	1	0.35	0.18
235	[Market equity FYE]/[Capex PPE sch V]	-1	0.35	0.30
236	[PPE land and improvement]/[Pension retirem expense]	-1	0.64	-0.00
237	[Interest expense]/[Cost of goods sold]	-1	0.35	0.63
238	[Operating expenses]/[Op income after deprec]	1	0.35	0.15
	Mean Data-Mined		0.43	0.44

References

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

Akerlof, G. A. and P. Michaillat (2018). Persistence of false paradigms in low-power sciences. *Proceedings of the National Academy of Sciences 115*(52), 13228–13233.

Back, K. (2010). *Asset pricing and portfolio choice theory*. Oxford University Press.

Bali, T. G., R. F. Engle, and S. Murray (2016). *Empirical asset pricing: The cross section of stock returns*. John Wiley & Sons.

Banz, R. W. (1981). The relationship between return and market value of common stocks. *Journal of financial economics 9*(1), 3–18.

Barberis, N. (2018). Psychology-based models of asset prices and trading volume. In *Handbook of behavioral economics: applications and foundations 1*, Volume 1, pp. 79–175. Elsevier.

Barry, C. B. and S. J. Brown (1984). Differential information and the small firm effect. *Journal of financial economics 13*(2), 283–294.

Berk, J. B., R. C. Green, and V. Naik (1999). Optimal investment, growth options, and security returns. *The Journal of finance 54*(5), 1553–1607.

Calluzzo, P., F. Moneta, and S. Topaloglu (2019). When anomalies are publicized broadly, do institutions trade accordingly? *Management Science 65*(10), 4555–4574.

Chen, A. Y. (2022). Most claimed statistical findings in cross-sectional return predictability are likely true. *arXiv preprint arXiv:2206.15365*.

Chen, A. Y. and M. Velikov (2022). Zeroing in on the expected returns of anomalies. *Journal of Financial and Quantitative Analysis*.

Chen, A. Y. and T. Zimmermann (2020). Publication bias and the cross-section of stock returns. *The Review of Asset Pricing Studies 10*(2), 249–289.

Chen, A. Y. and T. Zimmermann (2022a). Open source cross sectional asset pricing. *Critical Finance Review*.

Chen, A. Y. and T. Zimmermann (2022b). Publication bias in asset pricing research. *arXiv preprint arXiv:2209.13623*.

Chen, N.-F., R. Roll, and S. A. Ross (1986). Economic forces and the stock market. *Journal of business*, 383–403.

Chordia, T., A. Subrahmanyam, and Q. Tong (2014). Have capital market anomalies attenuated in the recent era of high liquidity and trading activity? *Journal of Accounting and Economics 58*(1), 41–58.

Cochrane, J. H. (2005). The risk and return of venture capital. *Journal of financial economics 75*(1), 3–52.

Cochrane, J. H. (2009). *Asset pricing: Revised edition*. Princeton university press.

Cochrane, J. H. (2017). Macro-finance. *Review of Finance 21*(3), 945–985.

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

Fama, E. F. and K. R. French (2010). Luck versus skill in the cross-section of mutual fund returns. *The journal of finance 65*(5), 1915–1947.

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

Gomes, J., L. Kogan, and L. Zhang (2003). Equilibrium cross section of returns. *Journal of Political Economy 111*(4), 693–732.

Goto, S. and T. Yamada (2022). False alpha and missed alpha: An out-of-sample mining expedition. *Working Paper*.

Green, J., J. R. Hand, and X. F. Zhang (2017). The characteristics that provide independent information about average us monthly stock returns. *The Review of Financial Studies 30*(12), 4389–4436.

Hartzmark, S. M. and D. H. Solomon (2013). The dividend month premium. *Journal of Financial Economics 109*(3), 640–660.

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

Harvey, C. R. and Y. Liu (2020). False (and missed) discoveries in financial economics. *The Journal of Finance 75*(5), 2503–2553.

Harvey, C. R., Y. Liu, and H. Zhu (2016). ... and the cross-section of expected returns. *The Review of Financial Studies 29*(1), 5–68.

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

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

Jegadeesh, N. and S. Titman (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of finance 48*(1), 65–91.

Jensen, M. C. and G. A. Benington (1970). Random walks and technical theories: Some additional evidence. *The Journal of Finance 25*(2), 469–482.

Jensen, T. I., B. T. Kelly, and L. H. Pedersen (2022). Is there a replication crisis in finance?

Kuhn, T. S. (1962). *The structure of scientific revolutions*, Volume 111. Chicago University of Chicago Press.

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

Mas-Colell, A. (1977). On the equilibrium price set of an exchange economy. *Journal of Mathematical Economics 4*(2), 117–126.

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

McLean, R. D., J. Pontiff, and C. Reilly (2020). Taking sides on return predictability. *Georgetown McDonough School of Business Research Paper* (3637649).

Merton, R. C. (1973). An intertemporal capital asset pricing model. *Econometrica: Journal of the Econometric Society*, 867–887.

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.

Shiller, R. J. (2003). From efficient markets theory to behavioral finance. *Journal of economic perspectives 17*(1), 83–104.

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

Sonnenschein, H. (1972). Market excess demand functions. *Econometrica: Journal of the Econometric Society*, 549–563.

Sullivan, R., A. Timmermann, and H. White (1999). Data-snooping, technical trading rule performance, and the bootstrap. *The journal of Finance 54*(5), 1647–1691.

Sullivan, R., A. Timmermann, and H. White (2001). Dangers of data mining: The case of calendar effects in stock returns. *Journal of Econometrics 105*(1), 249–286.

Treynor, J. L. (1962). Toward a theory of market value of risky assets. Final version in Asset Pricing and Portfolio Performance, 1999.

Yan, S. (2011). Jump risk, stock returns, and slope of implied volatility smile. *Journal of Financial Economics 99*(1), 216–233.

Yan, X. S. and L. Zheng (2017). Fundamental analysis and the cross-section of stock returns: A data-mining approach. *The Review of Financial Studies 30*(4), 1382–1423.

Zaffaroni, P. and G. Zhou (2022). Asset pricing: Cross-section predictability. *Available at SSRN 4111428*.

Zhang, L. (2005). The value premium. *The Journal of Finance 60*(1), 67–103.