The Pitfalls of Asset Management Research

Campbell R. Harvey*

Abstract

We now know that research findings favorable to the sponsor of the research should be discounted on the grounds of conflict of interest (e.g., tobacco companies or pharma companies). Incentives distort research findings. Is the same true in the field of finance? I argue that economic incentives distort outcomes in both academic and practitioner finance research. The problem is somewhat less severe in the practice of finance. An asset manager who overfits their backtest will likely underperform in live trading. As such, they will lose investors and suffer a damaged reputation. Asset management has no equivalent to academic tenure.

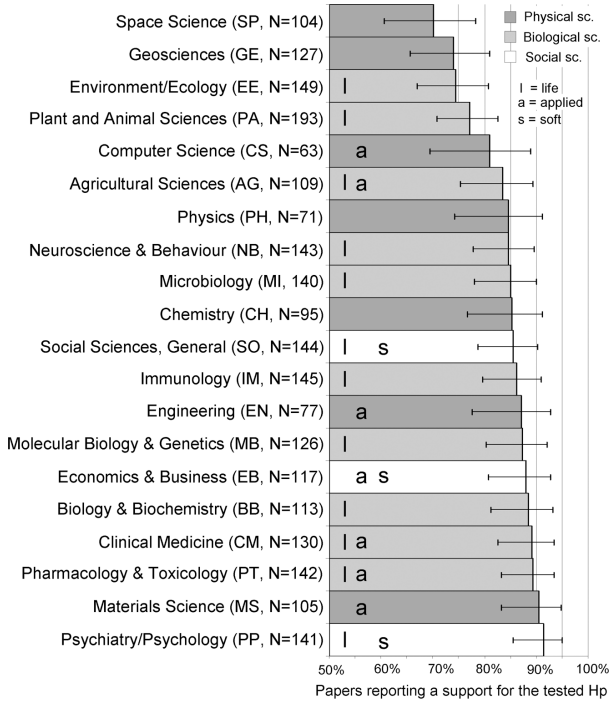
*. Duke University, Durham, NC 27708 USA National Bureau of Economic Research, Cambridge, MA 02138 USA.

Introduction

In my 35 years as an academic, as an advisor to many asset management companies, and as an editor of one of the top academic journals in finance, I now fully appreciate the crucial importance of the role incentives play in the production of research.

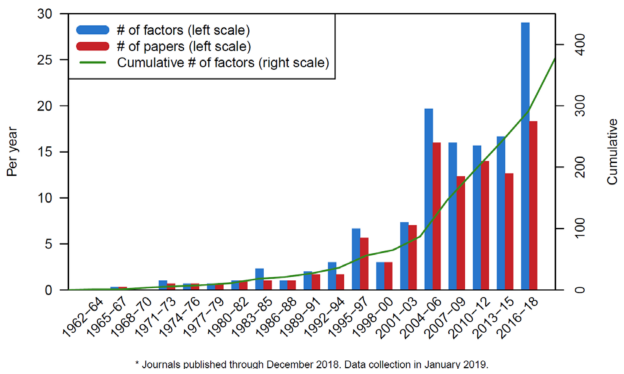
About 90% of the articles published in academic journals in the field of finance provide evidence in "support" of the hypothesis being tested.[1] Figure 1 shows the variation across different fields. Notice that space science and geosciences journals are far less likely to publish papers that provide "support" for the proposed hypothesis.

Figure 1: The proportion of papers that "support" the proposed hypothesis. Source: Fanelli (2010).



Indeed, my research detailed in Figure 2 shows that over 400 factors (strategies that are supposed to beat the market) have been published in top journals.[2] Almost all of these factors are "significant." How is that possible? Finding alpha is very difficult.

Figure 2: Factor production. Source: Harvey, Liu, and Zhu (2016) and Harvey and Liu (2020a).



Of course, many possible explanations exist some of which are discussed in Harvey (2017):

1. Our theories in financial economics are better than the theories in other fields. It is very difficult, however, to make the case that the theories in finance are better than, say, those in the field of particle physics.
2. We observe phenomena and hence are able to more precisely shape theories. This explanation has some merit. In some scientific fields, theories are proposed with no easy way to observe the data. It might be decades before a technology is developed to test the hypothesis. The discovery of the Higgs boson is a good example of this. It was proposed in 1964, the same year Sharpe published his initial paper on the capital asset pricing model (CAPM) for which he later won the Nobel Prize. Sharpe's theory was tested by 1972. Higgs had to wait decades after almost $5 billion was spent on the Large Hadron Collider in order to test his theory.
3. Our hypotheses are narrow and focused. This explanation also has some merit. A narrower implication can be more easily tested.
4. The connection between theories and empirical application are more flexible in finance. Again, this explanation makes sense. Think of testing the CAPM. We have many choices. What is the market portfolio? Do we test in

the United States or more broadly, such as globally? What time period do we test? How do we estimate the inputs?

5. In the field of financial economics, we may have more interaction effects between the researcher and the effect being tested. So-called confirmation bias impacts research in many social sciences, but is less likely to play a role in physical sciences.
6. Manipulation of the data and results may be more prevalent in finance. I will have more to say about this later with respect to *p*-hacking (strategic data selection, outlier exclusion, winsorization, data transformation, instrument selection, estimation methods, and so forth).
7. The lack of a replication culture in finance makes it easier to engage in *p*-hacking.
8. To publish a paper that does not find a "significant" result is difficult. As detailed in Figure 1, very few publications in finance present so-called negative results (i.e., results that do not "support" the null hypothesis).
9. The field of finance is more likely to ignore the multiple testing problem. Other fields have discovered that standard errors need to be inflated to control for multiple testing. That is, the *t*-statistic threshold needs to be higher for the 400th factor "discovery" than the first.

Incentives and publication

Consider the following narrative. Academic journals compete with impact factors, which measure the number of times an article in a particular journal is cited by others. Research with a "positive" result (evidence supportive of the hypothesis being tested) garners far more citations than a paper without significant or with negative results. Authors need to publish to be promoted (and tenured) and to be paid more. They realize they need to deliver positive results.

To obtain positive outcomes, researchers often resort to extensive data mining. While in principle nothing is wrong with data mining if done in a highly disciplined way, often data mining is not undertaken with discipline.

Researchers frequently achieve statistical significance (or a low *p*-value) by making choices. For example, many variables might be considered and the best ones cherry picked for reporting. Different sample starting dates might be considered to generate the highest level of significance. Certain influential episodes in the data, such as those arising from the global financial crisis or the COVID-19 pandemic, might be censored because they diminish the strength of the research results.

More generally, a wide range of choices for how to exclude outliers is possible as well as different winsorization rules. Variables might be transformed—for example, log levels, volatility scaling, and so forth—to get the best possible fit. The estimation method used is also a choice. For example, a researcher might find that a weighted least-squares model produces a "better" (more favorable) outcome than a regular regression.

The preceding options are just a sample of the possible choices researchers can make that all fall under the rubric *p*-hacking. Many of these research practices qualify as research misconduct, but are hard for editors, peer reviewers, readers, and investors to detect. For example, if a researcher tries 100 variables and only reports the one that works, that is research misconduct. If a reader knew the researcher tried 100 variables, they would also know that about five would appear to be "significant" purely by chance. Showing that a single variable works would not be viewed as a credible finding.

Multiple testing

To compound the problem, researchers do a poor job in controlling for luck. Suppose a researcher does not engage in *p*-hacking and fully reports that 100 variables were tried. The researcher claims five are significant, yet we know by random chance that five should appear to be significant.

Although statistical methods to account for the number of variables tried are readily available, they are rarely used.[3]

The incentive problem, along with the misapplication of statistical methods, leads to the unfortunate conclusion that roughly half of the empirical research findings in finance are likely false.

Incentives also differ across academic institutions. The very top schools do not just count the number of publications in tenure review cases. These schools also look for publications that will have a lasting impact. The paper that tries 100 variables and cherry picks the most significant one is unlikely to have a lasting impact, because the result is probably a fluke and further research will uncover its fragility. The vast majority of academic institutions, however, do simply count the *number* of publications for promotion decisions.

These circumstances suggest an important lesson: A peer-reviewed paper should be trusted more than a non-peer-reviewed paper, but skepticism of peer-reviewed papers is also warranted.

Is there a replication crisis in finance?

A robust debate is currently underway regarding a replication crisis in the field of finance. My coauthors and I started the debate several years ago when we stated: "A new factor needs to clear a much higher hurdle, with a *t*-statistic greater than 3.0. We argue that most claimed research findings in financial economics are likely false."[4] While our statement was provocative, many other fields have come to a similar realization. In particular, in 2005, regarding the field of medicine, Ioannidis asserted in his landmark paper that "most published research findings are false."[5]

Several years ago, my coauthors and I conducted a meta study and proposed various corrections to standard errors to take multiplicity into account.[6] A number of follow-up papers provided evidence. These papers include the following. McLean and Pontiff (2016) studied the out-of-sample performance of 97 factors. Out of sample is defined as post-publication. They found that returns are 26% lower out of sample and 58% lower post-publication. This evidence is consistent with in-sample overfitting. It is also consistent with some of these factors being arbitraged away after discovery.

Linnainmaa and Roberts (2018) presented an intriguing out-of-sample test using newly collected historical data (that factor researchers did not have previously). They stated "using data spanning the twentieth century, we show that the majority of accounting-based return anomalies, including investment, are most likely an artifact of data snooping."

Chen (2021) conducted an interesting thought experiment. Suppose all factors have a zero premium. How likely is it that we could observe the distribution of *t*-statistics documented in the literature? Chen's results suggest this is extremely unlikely, however, the results are difficult to interpret. The results suggest that not all factors are false, however, few researchers believe all factors are false.

Jensen, Kelly, and Pedersen (2022) presented an empirical Bayes model that challenges the need to inflate standard errors for multiple testing. To implement the model, the researcher must take a stand on the prior proportion of true findings. Harvey and Liu (2020a) showed the relation between the prior proportion (p_0 parameter in Figure 3) and the *t*-statistic cutoff. If a researcher believes that 50% are true, no inflation of standard errors is necessary.

External validation

While the replication debate rages, research papers continue to use academic factors. These factors ignore implementation costs, such as

Figure 3: The relation between prior belief and *t*-statistic using mutual fund returns. Source: Harvey and Liu (2020a).

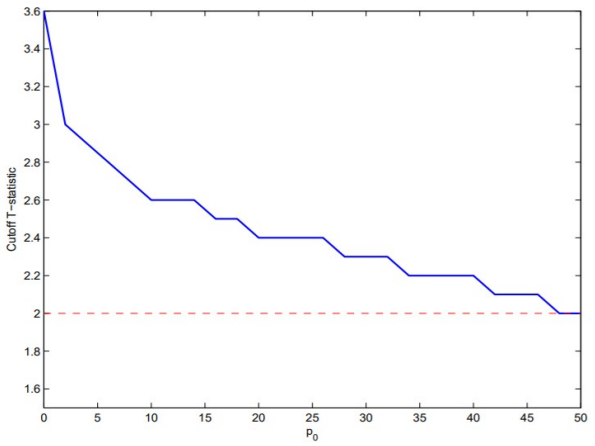
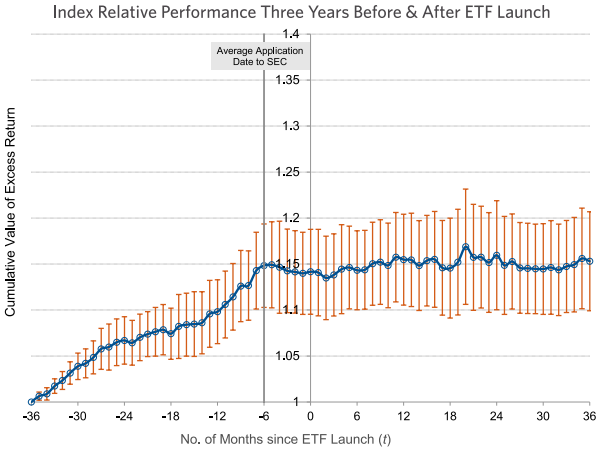


Figure 4: ETF performance after launch. Source: Research Affiliates, LLC, using data from Bloomberg.



transactions costs, borrowing costs for shorting, and so forth. These costs can be large.[7]

A potentially useful testing group is newly launched ETFs. Many new ETFs have claimed to be based on peer-reviewed research published in the finest academic journals. Few investors realize that peer-reviewed research could have been p-hacked or overfit to such an extent that the results are unlikely to repeat out of sample. Indeed, the evidence points very starkly to this phenomenon.[8]

Figure 4 illustrates the market-adjust (after all costs) returns of all ETFs. Notice that the backtested returns are strong. After application to the SEC and the subsequent launch of the ETF, however, the excess returns are zero. This outcome is consistent with overfitting and/or *p*-hacking.

We can draw a number of lessons from this exhibit. First, the exhibit is consistent with the idea that academic findings are overstated in the backtest. The second insight is subtler. In their discussion of the erosion of out-of-sample factor performance, McLean and Pontiff proposed two hypotheses: 1) the degradation could be due to overfitting and/or *p*-hacking, or 2) it could be due to returns being arbitraged away.[9]

Figure 4 is inconsistent with the latter explanation. If a strategy did produce alpha, an ETF is unlikely to be the first to harvest the alpha; a hedge fund would be much more likely to jump in first. Thus, we would expect the flattening of an ETF backtest *before* the application to the SEC. The exhibit does not show this. Indeed, the slope is very steep in the year before the application to the SEC. As a result, the evidence is consistent with backtest overfitting.

In their analysis, Brightman, Li, and Liu (BLL) used data through 2014. Recently, a new study by Ben-David, Franzoni, Kim, and Moussawi both updated the data used by BLL through 2021 and conducted new tests.[10] In their analysis, Ben-David et al. examined smart beta ETFs in the same type of event study that BLL used. This approach sharpens the focus to the particular ETFs that rely on academic factors.

The findings of Ben-David et al. are remarkably consistent with the findings of BLL. The smart beta ETFs do well in the backtest and have zero excess performance for periods of three and even five years. The Ben-David et al. paper provides both a

replication of BLL and a sharpened focus on smart beta ETFs.

Practitioner Research

Does the research conducted by asset management companies suffer from the same problems? The answer depends, again, on incentives.

Asset management companies engage in three categories of research. The first is research that is purely for client consumption and is published on company websites. The second is research that is meant for peer-reviewed practitioner or academic outlets. The final category is proprietary research that is not intended for publication, but could provide the foundation for product design. Although the research is not published, the details of the research will be shared with certain clients because they need to understand what they are investing in.

As with academic research, investors need to be skeptical of asset management research conducted by practitioners. Indeed, one company might comb through the academic research and do its own data mining in order to launch many ETFs, fully knowing some will fail. Nevertheless, the company receives a fixed fee. Given the large number of funds launched, most remember the winners more than the losers.

I often feature a paper from a well-known active asset manager that makes the case that active managers beat passive indices far more often than previously thought. The empirical results are obtained by excluding the bottom-performing quartile of active managers. Of course, 100% of active managers beat the S&P 500 Index if we censor from the sample every manager who underperformed! This is an example of strategic data selection (select only the data that support your hypothesis).

A company's research culture strongly influences the probability of *p*-hacking. Consider a company that has two high-quality researchers, let's call them A and B. Both of them pitch ideas to the CIO, and the CIO considers both research ideas equally promising. The research is conducted with great care and without *p*-hacking. A's idea shows considerable promise when applied to the data. B's idea fails. A's idea goes into live trading.

In my example, both A and B are equally high-quality researchers. The asset management firm makes a big mistake if A is promoted or given an extra bonus—or even worse, B is terminated. Such treatment leads to a dysfunctional research culture in which, at the beginning of a research project, both A and B realize they need to deliver statistically "significant" results in order to be promoted, or possibly to be retained, at the company. They may respond to these incentives by beginning to data mine and p-hack.

I believe *p*-hacking is less of a problem in asset management than in academia—in particular, less of a problem in the proprietary research that is the foundation for a product. The reasons are simple. First, in the presence of a performance fee, the asset management company's research needs to be optimized in a way that maximizes the chances of repeatable performance.[11] This means the asset manager does not choose the best-performing backtest, because it is the one that is most likely to be overfit. If the manager were to launch a backtest-overfitted strategy, it would likely fail and thereby generate no performance fees. The second reason is reputation. Academic tenure has no equivalent in asset management. If an asset manager's products disappoint because of overfitting, the firm's investors will flee. This market mechanism naturally minimizes the overfitting. That said, asset management companies still produce a substantial amount of low-quality research. Similar to the academic research, investors need to be skeptical.

Accounting for out-of-sample performance

A strategy might not perform as well as the backtest in live trading for a number of reasons, such as the following:

1. the in-sample performance was overfit (i.e., the "best" model was selected and is potentially overparameterized and p-hacked);
2. the researcher failed to account for multiplicity (i.e., what appears to be a significant predictor is not, so the strategy is destined to fail out of sample);
3. nonstationarity (i.e., the economic foundation of the strategy changes; for example, a regulatory loophole is closed and the alpha vanishes)[12]; and
4. bad luck (i.e., even though the strategy is a robust strategy, it may encounter a period of bad luck out of sample).

In live trading, investors need to decide whether to abandon the strategy. Reasons 1 to 3 point to abandoning the strategy because it should not work in the future. Reason 4 should be considered if no evidence exists to support reasons 1 to 3. Investors should be careful not to make a Type II error (false negative) by abandoning the strategy just before it turns around.

How can we improve?

The problem of *p*-hacking is not unique to finance. Indeed, many other academic areas are realizing they have the same problem and are taking steps to address it. Indeed, the cost of *p*-hacking is arguably much higher in medical research—where a study's results might mean the difference between life and death—than in finance where the primary concern is the size of the alpha. Currently, the field of finance is in the midst of a healthy debate about the severity of the replication crisis. Our field is not special compared to other fields of study and does not warrant a free pass.[13]

Investors can take a number of steps to mitigate the problem. First, be skeptical of both academic and practitioner research. Often a predetermined agenda or incentives make the results seem stronger than they are. Second, take the research culture into account. For example, when presented with a new strategy, ask if the company keeps a record of all variables that were tried. Third, try to quantify the costs of different mistakes (selecting a bad manager versus missing a good manager).[14] Fourth, make sure the strategy has a solid economic foundation. Also, beware of ex-post theorizing (after discovering the result, a story is concocted). Fifth, strategically ask questions such as "Did you try X?" If the answer is "Yes, and it does not work" and X is not reported, interpret this as a red flag. On seeing one cockroach, you can safely assume a dozen are behind the wall.

Notes

1. See Fanelli (2010) and Harvey (2017)
2. See Harvey, Liu, and Zhu (2016) and Harvey and Liu (2020a)
3. See Bailey and López de Prado (2014), Harvey and Liu (2014, 2015), and Chordia, Goyal, and Saretto (2020)
4. See Harvey, Liu, and Zhu (2016)
5. See Ioannidis (2005)
6. See Harvey, Liu, and Zhu (2016)
7. See Arnott et al. (2022)
8. See Brightman, Li, and Liu (2015)
9. See McLean and Pontiff (2016)
10. See Ben-David et al. (2022)
11. See Harvey and Liu (2018)

12. See Cornell (2022)

13. See Bailey and López de Prado (2021) and Harvey and Liu (2021)

14. See Harvey and Liu (2020b)

References

Arnott, Rob, Campbell R. Harvey, Vitali Kalesnik, Juhani T. Linnainmaa, and Lillian Wu. 2022. "Investible Factors." Work in progress.

Bailey, David H., and Marcos López de Prado. 2014. "The Deflated Sharpe Ratio: Correcting for Selection Bias, Backtest Overfitting and Non-Normality." *The Journal of Portfolio Management* 40 (5): 94–107.

———. 2021. "Finance Is Not Excused: Why Finance Should Not Flout Basic Principles of Statistics." *Royal Statistical Society (forthcoming).*

Ben-David, Itzhak, Francesco Franzoni, Byungwook Kim, and Rabih Moussawi. 2022. "Competition for Attention in the ETF Space." *Review of Financial Studies (forthcoming).*

Brightman, Chris, Feifei Li, and Xi Liu. 2015. "Chasing Performance with ETFs." *Research Affiliates (November).*

Chen, Andrew Y. 2021. "The Limits of *p*-Hacking: Some Thought Experiments." *The Journal of Finance* 76 (5): 2447–2480.

Chordia, Tarun, Amit Goyal, and Alessio Saretto. 2020. "Anomalies and False Rejections." *The Review of Financial Studies* 33 (5): 2134–2179.

Cornell, Bradford. 2022. "Data Mining, Nonstationarity, and Entropy: Investment Implications." *Journal of Systematic Investing* 2 (1).

Fanelli, Daniele. 2010. ""Positive" Results Increase Down the Hierarchy of the Sciences." *PLOS One* 5 (4): e10068.

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

———. 2022. "Be Skeptical of Asset Management Research." In *Investment Luminaries and Their Insights: 25 years of the Research Foundation Vertin Award,* edited by Bud Haslett, 35–38. Charlottesville, VA: CFA Institute Research Foundation.

Harvey, Campbell R., and Yan Liu. 2014. "Evaluating Trading Strategies." *The Journal of Portfolio Management* 40 (5): 108–118.

———. 2015. "Backtesting." *The Journal of Portfolio Management* 42 (1): 13–28.

———. 2018. "Detecting Repeatable Performance." *The Review of Financial Studies* 31 (7): 2499–2552.

———. 2020a. "A Census of the Factor Zoo." *Working paper (October 16). Available at SSRN.*

———. 2020b. "False (and Missed) Discoveries in Financial Economics." *The Journal of Finance* 75 (5): 2503–2553.

———. 2021. "Uncovering the Iceberg from Its Tip: A Model of Publication Bias and *p*-Hacking." *Working paper (June 30). Available at SSRN.*

Harvey, Campbell R., Yan Liu, and Heqing Zhu. 2016. ". . . and the Cross-Section of Expected Returns." *The Review of Financial Studies* 29 (1): 5–68.

Ioannidis, John P. A. 2005. "Why Most Published Research Findings Are False." *PLOS Medicine* 2 (8): e124.

Jensen, Theis Ingerslev, Bryan T. Kelly, and Lasse Heje Pedersen. 2022. "Is There a Replication Crisis in Finance?" *Journal of Finance (forthcoming).*

Linnainmaa, Juhani T., and Michael R. Roberts. 2018. “The History of the Cross-Section of Stock Returns.” *The Review of Financial Studies* 31 (7): 2606–2649.

McLean, R. David, and Jeffrey Pontiff. 2016. “Does Academic Research Destroy Stock Return Predictability?” *The Journal of Finance* 71 (1): 5–32.

Acknowledgments

April 4, 2022. Kay Jaitly provided editorial assistance. Some of material in the article is based on a chapter in Harvey (2022).