Does it take CRSP data to publish in the Journal of Finance? Analysis of database usage and publication trends

John Paul Broussard – Mika Vaihekoski*

December 8, 2025

Abstract

This paper studies the popularity of different databases in research articles published in the Journal of Finance. Analyzing 945 articles using three-year samples from 1992 onwards over four decades, we document that the CRSP is by far the most often used database, followed by Compustat (USA). However, their popularity has declined slightly in the 2020s amid the rise of non-US data sources. The results also show the strong impact of Professor French's freely accessible database on the academic community. More than six percent of the empirical articles in the sample have utilized this resource. At the same time, the results show that the average article length has increased by more than seven pages per decade. However, the growth rate has decreased, arguably partly because of the 60-page limit established in 2006 and the increased use of Internet Appendices in the 2010s. Finally, the results show that the number of authors has steadily increased from fewer than two in the 1990s to almost three in the present day.

JEL classification: A14, B26, G0.

Keywords: financial databases, Journal of Finance, CRSP, Compustat, SDC, IBES, Datastream, Bloomberg, Altmetric Score.

*In alphabetical order. Broussard: Rutgers University. Email: john.broussard@rutgers.edu. Vaihekoski (corresponding author): Turku School of Economics, University of Turku, Finland. Email: mika.vaihekoski@utu.fi. We are grateful for the helpful comments from Ville Kaukonen and other seminar participants at the 5th Turku Finance Research Seminar. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. The author reports there are no competing interests to declare.

High-quality databases play an essential role in research. Financial researchers often have the luxury of having data in high quantities, both cross-sectionally and historically. As such, the impact of high-quality databases on the phenomenal success of financial research comes as no surprise (Schwert, 2021). If one had to mention one database that has had probably the most significant impact on financial research, it is the databases on the US stock market created by the Center for Research in Security Prices, LLC (CRSP). CRSP was founded in 1960 by James H. Lorie and Lawrence Fisher of the University of Chicago. Its goal was to provide data that can be used to understand security markets in the USA.[1] The CRSP database still stands as the golden standard for research quality data on securities markets because of its high accuracy and comprehensiveness. Obviously, the fact that the US market is considered the most advanced in the world has also contributed to the widespread use of the database.

It is well known that access to novel, high-quality datasets can facilitate research success and potentially lead to higher-quality publications. This has led researchers to build new databases and constantly push the boundaries, inventing novel ways to use and connect existing and newly created databases. Technological development has also facilitated this process in many ways. For example, in the 2000s, we have seen several new big data databases emerge (c.f., e.g., Goldstein et al., 2024 and Schwert, 2021). Several researchers have also collected historical data to extend the temporal and global coverage of the data available for research (see, e.g., Dimson et al., 2002). At the same time, commercial vendors have recognized the importance of financial databases to the industry as a whole, which has also benefited researchers. History has proven there is a market for up-to-date data as well as historical data, even within the finance industry, and nowadays arguably even more than before, as artificial learning and different data centers require more and more data. As a result, researchers have begun using so-called alternative datasets that cover non-financial topics and are often collected for other purposes (see, e.g., Dessaint et al., 2024).

[1] https://www.crsp.org/crsp_pdf/history-james-lorie-benchmarks-for-research-quality-data/

While there are numerous bibliographic studies, e.g., on journal performance, co-authorships, university backgrounds, co-citation trends, as well as popularity of specific research streams within certain fields (a good review of financial research-related studies can be found in Khan et al., 2022), there are hardly any studies on what databases are used by the financial researchers. This is probably because such information is not readily available and must be manually collected for the most part.

This paper studies the usage of different databases in the *Journal of Finance*. We analyze more than 900 research articles published in three-year samples from 1992 onwards, one decade apart, to discover which databases are most frequently used and whether there are clear trends in their popularity. We also examine whether the data are mostly US-centric and whether other countries are also frequently analyzed. The data also allow us to study very specific questions, such as the role of the free-for-all data available on Professor Kenneth French's website. In addition, the collected data enable a detailed analysis of other important publishing trends in the Journal's research articles over the past 30 years. More specifically, we focus, among other things, on article length, co-authorship structures, as well as the impact of articles, both within and outside academia.

This article contributes to the literature in at least three respects. First, as far as we know, it is the first to conduct an empirical analysis of research data sources and database usage in one of the leading finance journals. Second, this article provides up-to-date information on publication trends that have taken place in the *Journal of Finance*. As such, it contributes to Borokhovich et al. (2000) on citation patterns and Pinkowitz (2002) on the variables influencing article downloads and their impact on citations. Third, the results of this article can be used as inputs for developing the *Journal* in the future if such a conclusion is warranted.

The rest of the paper proceeds as follows. Section I presents and describes the data used in our empirical analysis. Section II presents the main results from the empirical analysis. Section III provides a concluding discussion of our findings.

I. Data and Sample Selection

This study focuses on research articles published in *the Journal of Finance* (henceforth Journal), a journal known for its high quality and the variety of topics it covers in finance. For closer analysis, we select three-year subsamples ten years apart from 1992 onwards: 1992–1994, 2002–2004, 2012–2014, and 2022–2024. There are no specific reasons for selecting these years, other than the fact that this allows us to have as recent information as possible in the sample and then extend the analysis backward to find out more about the trends and potential evolving patterns in the results.

As the first step, we hand-collect background information on each article published during these years from the Journal's website: the article title, the authors' names, and the first and last page numbers. After that, we collect information on whether the article has an Internet Appendix, replication code available online, or both.[2] In addition, we collect information on whether the article is available as open access or free to read to the reader. Finally, we collect the number of citations and the Altmetric Attention Score for each article on Wiley's website.[3] The information is collected only for the research articles — including the shorter articles published in the 1990s and the presidential addresses. All non-research articles, including, e.g., front matter, book reviews, miscellanea, announcements, Minutes of the Annual Meeting, and associated reports, are excluded. Ultimately, our sample has 945 research articles.

The number of citations is often considered a measure of research quality. Altmetric score, on the other hand, is a widely used indicator of an article's impact outside academia, as it is based on the amount of attention an article has received online through news coverage, social media, government reports, and other sources. Altmetric score is not without its problems, and one has to remember that it is not a measure of the quality of the research (c.f., e.g., Ebrahimzadeh et al., 2022). However, despite its flaws, some universities and

[2] In the 2013 issue 2, eight articles had 'supporting info items'. We counted them as Internet Appendices.
[3] Citation data is collected during February 20–24, 2025, Altmetric data on October 31, 2025.

funders now use Altmetrics in research evaluation to demonstrate policy impact and media reach. Business schools have also begun citing them in accreditation and promotion reviews as proof of the relevance of the researcher's work beyond the academy.[4]

In the next step, the articles are analyzed to determine whether they are mainly theoretical or empirical. To be labelled as empirical articles, we require that real (non-simulated) data have been used meaningfully in the analysis. The rest are labeled as theoretical articles. There are some tough calls, but in general, the choice is rather straightforward. The main judgment call occurs when an article demonstrates the model developed in the article using a real-life setup or simulation, where starting values are potentially derived from real-life data. In most cases, these articles are marked down as theoretical unless empirical data has a role beyond setting the starting values for the analysis.

For the empirical articles, we collect information on all data sources used in each article. These sources are then divided into two groups: databases and other data (non-database) sources. The division allows us to separate databases from other data sources that are often unique and/or available only in written form. In addition, it facilitates our data collection, as for most databases, it is mostly a tick-the-box exercise. In practice, we begin with a list of several commonly used databases and expand it when we observe frequent use of databases that were not initially included. We also have a separate space to add those databases that appeared only a few times in the sample. For non-database sources, we record detailed information about the data and its sources for the article in question. These sources included, among others, data from other researchers, physical documents (e.g., books, reports), surveys, web-based data (e.g., LinkedIn, Google Trends), or proprietary data from a company.

There are, however, a large number of sources that are hard to categorize into either group. As the standard definition for databases requires that the data are stored and managed electronically, we labeled data that were manually collected and potentially digitized

[4]Murray (2025). https://www.ft.com/content/12aa378f-9c39-422b-802a-5b182c99a2bc?, accessed on November 3, 2025.

for an article, book, directory, or report as originating from a non-database source. A related question is the media used to access the data. We require that the database in question was available at the time of the study. A case in point is the Wall Street Journal (Index), which is now available online. However, as online access to the WSJ began in the mid-1990s, data for the articles published in the 1990s in our sample must have been collected manually from physical copies. This could also be true for newer articles, but they often remain silent about the media used to access the content. As a result, WSJ and other similar newspapers and magazines are categorized as non-database sources unless the article clearly notes using an online database to access its content. The same approach is used for all similar sources. A good example is Cooperman et al. (1992), which uses eight different non-database data sources (ranging from a survey to the WSJ Index and regional newspapers).

Another group of hard-to-decide cases is those articles that mention only a company as the source for their data (e.g., Credit Suisse, MSCI, Zacks Investment Research, or a stock exchange). In these cases, we are quite liberal and label almost all of them as database use unless it is obvious from the text that it is not the case. There are also articles that do not mention a specific source for the data. There might be several reasons for this, ranging from lax review, especially earlier in the sample, to verbally or collectively shared information regarding the source. The authors might also have considered the source so well known that they decided not to mention it (or forgot to). As an example, one can mention Campbell and Hamao (1992), which uses the Japanese call money rate and the Gensaki rate, but no source for the data is given. A more recent example is Ramadorai (2012), which uses the VIX index (among others), but again, no source for the data is given. As there are typically several sources for the same data, some of which are of higher or lower quality, citing the source would help the reviewer evaluate the legitimacy and replicability of the results.[5]

A totally different issue in the data collection is that the article may note that the data were obtained, e.g., from Wharton Research Data Services (WRDS) or Datastream, but

[5]Both examples are randomly selected examples of similar cases.

the database used is different. For example, WRDS is a platform that provides access to databases from multiple vendors – more than 600, as of now.[6] WRDS does provide access to some databases of its own. Datastream is another type of platform that one can typically use to access a wide variety of contemporary and historical financial data. Again, some of the available data might be originally from another data vendor's database, some are originally from the London Stock Exchange Group – the company behind Datastream. Here, in both cases, we registered the original database if given, but if the (underlying) database was not mentioned, we registered the platform used to access the data.

The final, minor issue was due to the fact that the names of the databases or the platforms used to access the data have changed over time. This was especially the case with Datastream and related products. Namely, Datastream was originally developed by ICV (International Computation of Values) in the 1960s. It later became part of Primark Corporation, which was acquired by Thomson Financial in 2000. In 2008, Thomson Corporation merged with Reuters to form Thomson Reuters. In 2018, Blackstone Group acquired a 55% stake in Thomson Reuters' financial and risk business, creating Refinitiv, which subsequently acquired Datastream. Then in 2021, London Stock Exchange Group (LSEG) acquired Refinitiv. As a result of the company name changes, access to Datastream has been available, e.g., via Reuters 3000 Xtra (developed in the late 1990s), Thomson ONE (developed in the early 2000s), Refinitiv Eikon (developed in 2010), or LSEG Workspace (developed in 2021).[7] Luckily, researchers have most often named Datastream as their source, as it was often more convenient for accessing the data. Here, all references to Datastream and the related products are counted as one and the same.

A wide range of databases has been used in the articles. The nine most commonly used databases are referred to here by the following acronyms: CRSP, Compustat, Datastream, Bloomberg, Prof. French, EDGAR, SDC, IBES, and FRED. CRSP refers to the historical databases that the Center for Research in Security Prices has collected and created. They

[6] https://wrds-www.wharton.upenn.edu/ accessed February 15, 2025.
[7] Information on the products is collected from the LSEG website www.lseg.com.

are best known for daily US stock market price and index data going back to the early 20th century, but several articles in our sample also used their other databases (e.g., the bias-free mutual fund database). Compustat is a database of financial, statistical, and market information on active and inactive companies worldwide. The service was founded (and is still owned) by Standard and Poor's in 1962. It provides access to standardized financial statements from both US and global companies.

Datastream refers to its namesake platform for accessing data, as well as to other LSEG (and earlier companies') platforms that provide information primarily on publicly traded companies and assets worldwide. Bloomberg is a financial information terminal that provides access to a wide range of real-time and historical information on publicly traded assets and companies, as well as markets worldwide. The first version of the terminal was released in 1982.

SDC refers to the SDC Platinum database. It was originally developed by the Securities Data Company in 1969, but is now also owned by the London Stock Exchange Group (LSEG). The SDC database provides comprehensive data on financial transactions, including mergers and acquisitions, new issues, and other relevant information. In December 2023, the legacy SDC Platinum desktop application was discontinued, and its functionalities were integrated into LSEG's modern platforms, such as Workspace.[8]

IBES refers to the I/B/E/S (Institutional Brokers' Estimate System) database. I/B/E/S provides financial estimates and analyst forecasts, widely used for earnings analysis and investment research. The service was founded in 1976 by Lynch, Jones & Ryan, and Technimetrics, Inc., and has undergone several ownership changes over the years. In 1993, it was acquired by Barra, then sold to Primark Corporation in 1995, which was later acquired by Thomson Financial. At the moment, it is part of the LSEG group.

Prof. French refers to the data that Professors Eugene Fama and Kenneth French have made available on Professor Kenneth French's website, or earlier, upon request, directly from

[8] https://www.lseg.com/en/data-analytics/products/sdc-platinum-financial-securities, accessed on February 15, 2025.

the authors, as shown in Dittmar (2002). We are not aware of when the website was set up, but the earliest reference to it in the Journal is in Kumar and Lee (2006). EDGAR refers to the Electronic Data Gathering, Analysis, and Retrieval system set up by the U.S. Securities and Exchange Commission (SEC). It was established as a pilot program in 1984 and opened to the public in 1992. Since May 1996, companies have been required to file all material electronically. Nowadays, it is accessible as an online database that provides, among other things, access to companies' registration statements, prospectuses, and periodic reports.[9] FRED refers to the Federal Reserve Economic Data – St. Louis Fed's signature online database. It was established in 1991 as a free electronic bulletin board, and in 1995 it moved to the internet (Federal Reserve Bank of St. Luis, 2014).

A more detailed explanation of these databases can be found in the Appendix. We have also included rough guidance for the annual cost of university (research) access to the data(base) using the most inexpensive option. For example, in the case of Bloomberg, the cost refers to one Bloomberg terminal. Note that prices may be higher for companies. For commercial products, discounts may be available for larger purchases, but these details are not provided.

The articles also use a large number of other databases. We also collected the names of these databases. In this group, there are several commercial databases (e.g., Orbis and Crunchbase), but also government databases (e.g., OECD National Accounts database), databases created by organizations (e.g., Berkeley Options database), and, as said, data sources referred to by the company name (e.g., Iterative Data Services ISL). However, their usage was rather limited compared to the nine databases selected for closer analysis.

Finally, for empirical articles, we also mark down whether the data used in the article is focused on the US market, the global market, a specific region (e.g., East Asia or Europe), or a single country. There are also several instances in which data from two countries (e.g., the US and Canada) are used in the article. Deciding on the country required some subjective

[9] https://www.sec.gov/info/edgar/regoverview.htm, accessed on October 28, 2025.

evaluation, even though, in most cases, choosing the focus country (or area) is relatively easy. The main issues took place with articles analyzing doctoral graduates (Zivney and Bertin, 1992), FX markets (see, e.g., Biais et al., 2023), or data from a lab survey (see, e.g., Charles et al., 2024) or FinTech App (Gargano and Rossi, 2024). If the responses were collected from US respondents or the analysis was done from the US investors' point of view (e.g., FX or mutual fund returns are measured in USD), we labeled the data to focus on the US. While collecting the data, we observed that the US market was often assumed as the default, and it was not explicitly mentioned in the majority of the articles.

II. Results

A. Publication trends

The average number of articles published each year is 78.8. The number has remained relatively stable in all subsamples, as shown in Figure 1. In the 1990s, the Journal published five issues each year, but starting from volume 53 in 1998, a sixth issue was added. This does not seem to have increased the number of articles published in a year, as the length of the articles increased at the same time. In the 2010s, the number of published articles per year actually decreased quite dramatically, but one can see a slight increase towards the end of the sample.

[Insert Figure 1 near here]

It is well known that articles in top journals have become longer (see, e.g., Spiegel, 2012). Spiegel shows that the average length of the articles published in the Journal in 1980 was a mere 11.411 pages (with a median of 11 pages). Ten years later, it had grown to 17.511 pages (median 17), and another ten years later, in 2000, it was already 31.202 pages (median 30). In 2010, growth had slowed a bit, but the average length still increased to 34.275 pages (median 34). Here, we can confirm the growth using larger samples and show that it has

continued even in the 2020s, as we can also see from Figure 1. The results in Table I show that the average article length in the 2020s is 46.85 pages, with a median of 47.

In relative terms, growth has slowed. In the 2020s, the average article length grew by 20.4% from the decade earlier, whereas in the 2010s, the growth was clearly higher, at 27.1%. This could be because the *Journal* set a 60-page limit for articles in 2006. An indirect effect could be that more and more articles include Internet Appendices (IAs). Namely, even though the IAs are included in the 60-page limit, the authors may use them to report less important results, some of which may have been included in the original articles before the IAs became popular.

One of the earliest articles to make a specific reference to an Internet Appendix is Dewenter and Malatesta (1997) as far as we know. The publisher's (Wiley) webpage does not include Internet Appendices prior to 2012. Before that, the articles typically noted that an Internet Appendix is available in the Supplements and Datasets (section at `http://www.afajof.org/supplements.asp`). After a modest start, the popularity of Internet Appendices started to grow in 2009. In our sample, there are 100 articles (50.5% of all articles) with an IA published in the 2010s, and 185 (80.1%) in the 2020s sample (c.f., Table I). It has become quite uncommon to see an article without the IA. In 2016, the *Journal* implemented a code-sharing requirement, requiring authors of published papers to attach the program code used to generate the results with their articles. As a result, from 2018 onwards, more and more articles began to include the code. In our sample, 191 articles have code attached (82.7% of articles published in 2022-24).

There are likely to be many reasons for longer articles. Namely, the finance profession as a whole has developed over time, and the best articles are nowadays meticulously refereed from multiple angles, especially for the top journals. This often leads to increasing robustness analyses to answer all of the reviewers' concerns (Hirshleifer, 2014). Second, the articles have more authors, which can lead to longer articles, as each can add something to the paper.[10]

[10]Obviously, the causal relationship can also run in the opposite direction. Longer articles or more technical review reports may require more (specialized) authors to keep the amount of work per person manageable.

Third, multiple co-authors make it also easier to present the paper in many places, as the workload is shared. Each presentation is likely to lead to further improvements to the paper, which often translate into a longer article.

We run a simple regression to test which factors influence article length. In line with the overall development of the profession, we include the number of years since publication in the model. As articles labeled by the *Journal* as Shorter Papers are by definition shorter, we add an indicator variable into the model for them. In addition, we have a variable for the number of authors and a dummy for the Internet Appendices. Since there is reason to believe that empirical papers are longer than theoretical papers, we also include a dummy variable for them. Ultimately, we run the following regression:

$$\begin{aligned} No\_Pages_i = a_0 + a_1 Years\_SP_i + a_2 No\_Authors_i \\ + a_3 No\_Data_i + a_4 IA_i + a_5 SP_i + \epsilon_i, \end{aligned} \tag{1}$$

where No_Pages_i is the lenght of the article i (number of journal pages), $Years\_SP_i$ is the number of years since the publication (i.e., 2025 – the year of publication), $No\_Authors_i$ is the number of authors, No_data_i is an indicator variable with value one if no empirical data was used in the article (i.e., a theoretical paper), IA_i is an indicator variable with value one if the article has an Internet Appendix, and SP_i is an indicator variable for the shorter papers.

As our dependent variable takes only integer values, we estimate the equation as a count model. Now, comparing different estimation methods (e.g., Poisson model), the best fit is achieved under normal distribution (adjusted R-squared is 59.8%). The model is estimated using quasi-maximum likelihood (QML), which yields consistent parameter estimates even when the distribution is incorrectly specified. The standard errors are adjusted for heteroscedasticity using the Huber-White method adjusted for the degrees of freedom.

The results (available upon request) show that older articles are expectedly shorter ($a_1 = -0.019$, t-value = –15.463). Similarly, shorter papers are indeed shorter ($a_5 = -0.573$, t-value = –15.676). A higher number of authors increases the length of the article ($a_2 = 0.018$, t-value = 2.060). The other two indicator variables, No_Data_i and IA_i, are not found to influence the length of the article. Note that one cannot interpret the coefficients as one could for the least squares regression.[11]

We can analyze how the average number of coauthors has developed over time. It is well known that the number of co-authors has increased in top economics journals. Card and DellaVigna (2013) note that the number of authors per paper has increased from 1.3 in 1970 to 2.3 in 2012, partly offsetting the fall in the number of articles per year. There are reasons to believe that this trend has continued. Namely, there is arguably increased pressure to publish in top journals, from both a personal (tenure-track) and an employer's (university rankings) perspective. Furthermore, the articles require more multidisciplinary knowledge, e.g., by combining financial knowledge with econometrics, data science, and programming, which can lead to a higher number of co-authors. Larger workloads to meet the demand of top journals can also lead to more cooperation than before. Finally, Card and DellaVigna (2013) note that citation counts are significantly higher for longer papers and those written by more coauthors.

Spiegel (2012) shows that for the *Journal*, the mean number of co-authors has increased from 1.632 in 1980 to 2.418 in 2010. The same kind of increase can be seen in our subsamples, and the rise has continued in the 2020s. Table I shows that the average number of co-authors in the last subsample is 2.77, up from 2.47 in the 2010s.[12] The development for the *Journal* can also be seen in Figure 2.

[11] We also estimate the model using OLS, as the interpretation of the coefficients is easier. The results are essentially similar, but now the number of authors is no longer significant at the 10 percent level. The results show that shorter papers are, on average, 11.079 pages shorter, and that articles grow by 0.646 pages each year. We also test a model that includes dummies for the lead articles and presidential addresses. Neither variable is significant.

[12] The article published in 2024 with more than 100 co-authors has been excluded from the analysis as an outlier.

[Insert Figure 2 near here]

Next, we focus on the number of citations and the Altmetric attention score. Figure 3 shows the average number of citations to and altmetrics score for each research article published in the given year. As expected, older articles have more citations because they have had more time to accrue them. Altmetrics scores, on the other hand, behave differently: the average score decreases as one goes further back in history. This is arguably because the score cannot track contemporary attention for articles published in the 1990s, as Altmetric was created in 2011 (Nuredini, 2021). On the other hand, social media activity has increased over time, and as a result, references to contemporary research output have increased, leading to a higher score. The highest yearly average score is for the articles published in 2022, but among our three-year samples, the highest average (14.48) and median (10.00) Altmetric scores are for the articles published between 2012-14. The highest single-article score, 456, is for the Santa-Clara and Valkanov (2003) article, which studied the relationship between political cycles and the stock market, dubbed the Presidential Puzzle.

[Insert Figure 3 near here]

We can also test what influences the number of citations. The natural explanatory variable is the article's age, but as citations arguably increase exponentially, we also need to account for nonlinearity. There is also some evidence that longer text and more co-authors lead to more citations (c.f., Card and DellaVigna, 2013). There are also reasons to believe that empirical papers may differ from theoretical papers in either direction in their usage as references. Finally, shorter papers and those behind the paywall might get fewer citations, while those with an Internet Appendix might get more. Thus, we run the following cross-sectional regression

$$No\_Cites_i = a_0 + a_1 Years\_SP_i + a_2 (Years\_SP_i)^2 + a_3 No\_Authors_i + a_4 No\_Pages_i + a_5 No\_Data_i + a_6 IA_i + a_7 SP_i + a_8 OA_i \epsilon_i, \quad (2)$$

where No_Cites_i is the number of citations the article i has collected so far, $Years\_SP_i$ is the number of years since the publication, $No\_Authors_i$ is the number of authors, No_Pages_i is the number of pages, No_data_i is a dummy with value one if no empirical data was used in the article (i.e. a theoretical paper), IA_i is a dummy with value one if the article has an Internet Appendix, OA_i is a dummy with value one if the article is available as open access or with free to read status, and SP_i is a dummy for the shorter papers. We also test a model where we add two new dummies: $Pres\_Address_i$ and $First\_article_i$. The former variable takes the value one if the article is the presidential address, and the latter if the article is the lead article in an issue. Both of these article types are expected to get more citations.

The results from the OLS estimations are shown as Models 1 and 2 in Table II. Since the number of citations is always a positive integer, we also estimate a count model where we assume a Poisson distribution for the number of citations. In addition, as the relationship between the age of the article and the number of citations is likely to be highly nonlinear, we also test a model using the OLS, where we use the natural logarithm of $1 + No\_Cites_i$ as the dependent variable. The results are shown as Models 3 and 4 in Table II. Heteroscedastic-consistent standard errors have been used in all of the estimations.

It is clear from the results that the number of citations decreases with time (i.e., newer articles have fewer citations) but not linearly. Consistent with prior studies, the number of authors appears to be associated with more citations. This can be explained by larger combined networks of multiple authors, as well as better opportunities to present the paper across various forums, due to shared workloads. Longer papers get slightly more citations, and articles designated as shorter get clearly fewer citations. Theoretical papers also get

fewer citations than empirical papers. Having an Internet Appendix and making the article freely accessible both increase citations. Lead articles get more citations, but presidential addresses (also lead articles) do not get any special increase in their number of citations.

Finally, we test whether the same model can also explain the Altmetric Attention Score (AAS). The results are reported in Table II. Again, using a reduced-variable version of equation (2), we can see that the results differ clearly from those for the citation counts. Especially, the model's explanatory power is much lower (2.9% vs. 15.8%). This is driven by the fact that the score is highly skewed (9.29) with 26% of papers having an AAS value of zero. Thus, we apply a log transformation to the score (plus 1), reducing skewness to 0.26. Re-estimating the equation with two additional variables, as in Model 4 in Table II, we observe that the model's explanatory power is now clearly higher (20.7%). Seven variables are significant. Again, the time since publication is significant, with evidence of nonlinearity. Theoretical articles get predictably a clearly lower altmetric score (-0.296, t-value = -3.252), whereas open-access articles get expectedly a higher score (0.501, t-value = 6.566). The length of the article is also associated with a higher score (0.024, t-value = 5.169). In addition, articles with internet appendices are found to receive higher scores. The lead articles also receive a higher attention score (0.334, t-value = 2.042). However, as the model's explanatory power is lower than for citations (adjusted R-squared of 20.7% vs. 71.9%), an alternative approach to explaining the Altmetrics score is proposed, based on the idea that the article title may contribute to its attention.

To this end, we test whether the attention score is related to the article title using natural language processing (NLP) tools. In practice, we use word frequency analysis, Latent Dirichlet Allocation (LDA) topic modeling, Term Frequency – Inverse Document Frequency (TF-IDF) feature extraction with Lasso regularization, and multivariate regression. We find robust, statistically significant evidence that title content predicts attention, even controlling for the same variables as before. The results for the controlling variables are reported as Model 3 in Table II. We can see that the explanatory power has increased 4.5 percentage

points to 25.1%. The word frequency analysis found five clearly significant (at the 5% level) and eleven marginally significant (at the 10% level) terms. Papers emphasizing political economy, banking, and policy issues receive significantly more attention than those focusing on technical topics related to derivatives and abstract theory. Based on the results, if one wants to increase the impact of the research, it helps to frame the paper around policy implications, connect them to contemporary debates and certain institutes, and emphasize empirical contributions.

B. Database usage

As a starting point for the database usage analysis, we count the number of theoretical articles that do not use any data. These articles typically develop a theoretical model to analyze a real-life situation. Some articles also use simulation analysis to show the implications of the model given certain real-life derived parameter values. All in all, there are 188 (19.9% of all) theoretical articles (see Table IV) and 757 empirical articles. Theoretical papers were rather common in the 1990s (29.5% of all), after which their share has dropped to a slightly lower, yet quite stable, level.

Next, we count the number of times one or more different data sources are used in the empirical articles. Altogether, we identify 2,511 usages of different data sources. Of these, 2,227 are labeled as databases and 284 as other data sources. Across 757 empirical articles, we see that, on average, 3.31 different data sources were used per article. The results in Table IV show that the popularity of using more than four different data sources has increased over time. As an example of a recent article that used multiple data sources, see Allen et al. (2024). Namely, they use eleven different sources of data in their article (CRSP, Compustat, Datastream (Worldscope), BoardEx, RisMetrics, The Chinese Industrial Enterprises Database CIED, WIND, Chinese Capital Market Research Group CSMAR, National Bureau of Statistics, World Bank, and World Federation of Exchanges). In this case, all used data sources are also categorized as databases.

Analyzing the popularity of the nine most commonly used databases, we find that they are used a total of 893 times in our sample. As shown in the last column of Table IV, the most popular database during the sample period was, as expected, CRSP. It was used in 46.0% of the empirical articles. Thus, the answer to the outset of this article is clearly no – CRSP data is not a precondition for publication, but surely it does not hurt either.

The second most popular was Compustat (both US and Global databases), used in 30.5% of the articles. Somewhat unexpectedly, SDC was the third most popular with 8.5% usage. Its usage has almost diminished in the 2020s, though. Datastream (including other LSEG databases) has gained popularity, ranking fourth among databases with 7.4%.

Somewhat surprisingly, the data on Professor French's website ranked fifth among the most popular databases, even though it was not available in the 1990s. It was used in 6.3% of the articles. In the 2020s, the data were used in over ten percent of articles. It is clear that this database has had an enormous impact on the profession, enabling many studies that would otherwise have been undone.

The following four most popular databases – IBES, Bloomberg, EDGAR, and FRED – were each used in fewer than six percent of the articles. Obviously, many other databases were also used by the researchers. This category includes a variety of databases ranging from commercial databases (e.g., Orbis, NYSE TAB, OptionMetrics, Preqin, Morningstar, Factiva) to databases with an academic background (e.g., WRDS, CSMAR) and certain central bank/governmental databases.

We also analyze the popularity of the five most commonly used databases across time. In Figure 4, we can see that, somewhat unexpectedly, the popularity of the top-three databases has decreased in the 2020s. Especially, the usage of CRSP and Compustat has halved since the 2010s. This could be due to new alternative datasets that have become available more recently. In addition, finding new, meaningful insights from databases studied by hundreds of researchers becomes increasingly difficult.

[Insert Figure 4 near here]

C. Country analysis

Finally, we analyze the sample country (or area) of focus in the empirical articles. If an article covered data from several countries, both countries/areas (e.g., the US and Canada) are recorded for the analysis. All in all, the 757 empirical articles covered 764 countries or specific areas. Altogether, data from 22 different countries are analyzed in the articles and three different areas (global markets, Europe, and East Asia).

The results show that in 610 empirical articles (79.7%), the sample data is for the US market (see Table IV). Global markets are covered in 11.0% of the articles. Other markets are the focus of 9.3% of the articles. There are arguably several reasons behind the US-centricity in the empirical analysis. Namely, US financial markets are still, in many ways, considered the most developed ones and, as such, a natural starting point and benchmark for empirical analysis. Second, the availability of high-quality databases for the US market and the lack of comparable databases in other markets can influence the choice of sample data. Obviously, behavioral factors can also be at play. Based on prior publication patterns, authors can infer that US data can facilitate publishing decisions. Finally, one has to remember that journals always reflect the backgrounds of their readership and editors, for better or worse.

However, the usage of US data from outside the US market has increased almost linearly over time. In the 1990s, 90.2% of articles used US data, whereas in our 2020 sample, 72.2% did. The increase can be attributed to some degree to the use of global data (from nine to fourteen articles, an increase of 2.3 percentage points) and to the increased use of Chinese data (from zero to twelve articles, an increase of 1.6 percentage points). However, the most considerable increase comes in the use of data from countries or areas other than these. Namely, in 2022–24, 20.3% of articles used data from another country. These countries or areas include, among others, the UK, Europe, Japan, Canada, and Germany, to name a few. A complete listing, including a frequency analysis, is shown in Figure 5.

[Insert Figure 5 near here]

III. Conclusion

In this paper, we examine a novel topic in finance research: the use of diverse data sources and databases in research articles in the *Journal of Finance* over the last four decades. The results show that the most popular database during the sample period was, expectedly, CRSP. It was used in 46.0% of the articles. The second most popular was Compustat, used in 30.5% of the articles. Their popularity, however, has almost halved in the 2020s compared to earlier periods. This may reflect the emergence of new databases and the saturation of research on them, as hundreds of researchers have studied these databases since their introduction in the 1960s. The analysis also reveals that published empirical articles have primarily used US data, but data from other countries and regions are increasingly used, especially in the 2020s.

In addition, we examine several publication trends in the *Journal* and report several notable findings. The length of articles has continued to increase in the 2020s, despite a 60-page limit set in 2006. Growth has slowed somewhat in relative terms. This could arguably be due to the increased use of Internet Appendices, which are included in more than 80 percent of articles in the 2020s.

Moreover, consistent with prior studies, we find that the number of citations increases nonlinearly over time. The number of coauthors has a positive effect on citation count. Citations are also higher if the article has an Internet Appendix, is freely available to the public, and is the lead article in an issue. Theoretical articles and articles labeled as Shorter Papers by the *Journal* get fewer citations. The same model can be used to some degree to explain an article's Altmetric Attention score, but adding textual analysis improves the fit and reveals that if one wants to increase the impact of the research, it helps to frame the paper around policy implications, connect them to contemporary debates and certain institutes, and emphasize empirical contributions.

REFERENCES

Allen, F., Qian, J. Q., Shan, C., and Zhu, J. L. (2024). Dissecting the long-term performance of the chinese stock market. *The Journal of Finance*, 79(2):993–1054.

Biais, B., Bisière, C., Bouvard, M., Casamatta, C., and Menkveld, A. J. (2023). Equilibrium bitcoin pricing. *The Journal of Finance*, 78(2):967–1014.

Borokhovich, K. A., Bricker, R. J., and Simkins, B. J. (2000). An analysis of finance journal impact factors. *The Journal of Finance*, 55(3):1457–1469.

Campbell, J. Y. and Hamao, Y. (1992). Predictable stock returns in the united states and japan: A study of long-term capital market integration. *The Journal of Finance*, 47(1):43–69.

Card, D. and DellaVigna, S. (2013). Nine facts about top journals in economics. *Journal of Economic Literature*, 51(1):144–61.

Charles, C., Frydman, C., and Kilic, M. (2024). Insensitive investors. *The Journal of Finance*, 79(4):2473–2503.

Cooperman, E. S., Lee, W. B., and Wolfe, G. A. (1992). The 1985 ohio thrift crisis, the fslic's solvency, and rate contagion for retail cds. *The Journal of Finance*, 47(3):919–941.

Dessaint, O., Foucalt, T., and Fresard, L. (2024). Does alternative data improve financial forecasting? The horizon effect. *The Journal of Finance*, 79(3):2237–2287.

Dewenter, K. L. and Malatesta, P. H. (1997). Public offerings of state-owned and privately-owned enterprises: An international comparison. *The Journal of Finance*, 52(4):1659–1679.

Dimson, E., Marsh, P., and Staunton, M. (2002). *Triumph of the Optimists: 101 Years of Global Investment Returns*. New Jersey: Princeton University Press.

Dittmar, R. F. (2002). Nonlinear pricing kernels, kurtosis preference, and evidence from the cross section of equity returns. *The Journal of Finance*, 57(1):369–403.

Ebrahimzadeh, S., Alperin, J. P., and Haustein, S. (2022). Chapter 13: Social media and altmetrics. In *Handbook on Research Assessment in the Social Sciences*. Edward Elgar Publishing, Cheltenham, UK.

Federal Reserve Bank of St. Luis (2014). 100 years of service, 2013 annual report.

Gargano, A. and Rossi, A. G. (2024). Goal setting and saving in the fintech era. *The Journal of Finance*, 79(3):1931–1976.

Goldstein, I., Spatt, C. S., and Ye, M. (2024). The next chapter of big data in finance. *The Review of Financial Studies*, page hhae083.

Hirshleifer, D. (2014). Editorial: Cosmetic surgery in the academic review process. *The Review of Financial Studies*, 28(3):637–649.

Khan, A., Goodell, J. W., Hassan, M. K., and Paltrinieri, A. (2022). A bibliometric review of finance bibliometric papers. *Finance Research Letters*, 47:102520.

Kumar, A. and Lee, C. M. (2006). Retail investor sentiment and return comovements. *The Journal of Finance*, 61(5):2451–2486.

Nuredini, K. (2021). Investigating altmetric information for the top 1000 journals from handelsblatt ranking in economic and business studies. *Journal of Economic Surveys*, 35(5):1315–1343.

Pinkowitz, L. (2002). Research dissemination and impact: Evidence from web site downloads. *The Journal of Finance*, 57(1):485–499.

Ramadorai, T. (2012). The secondary market for hedge funds and the closed hedge fund premium. *The Journal of Finance*, 67(2):479–512.

Santa-Clara, P. and Valkanov, R. (2003). The presidential puzzle: Political cycles and the stock market. *The Journal of Finance*, 58(5):1841–1872.

Schwert, G. W. (2021). The remarkable growth in financial economics, 1974–2020. *Journal of Financial Economics*, 140(3):1008–1046.

Spiegel, M. (2012). Reviewing less–progressing more. *The Review of Financial Studies*, 25(5):1331–1338.

Zivney, T. L. and Bertin, W. J. (1992). Publish or perish: What the competition is really doing. *The Journal of Finance*, 47(1):295–329.

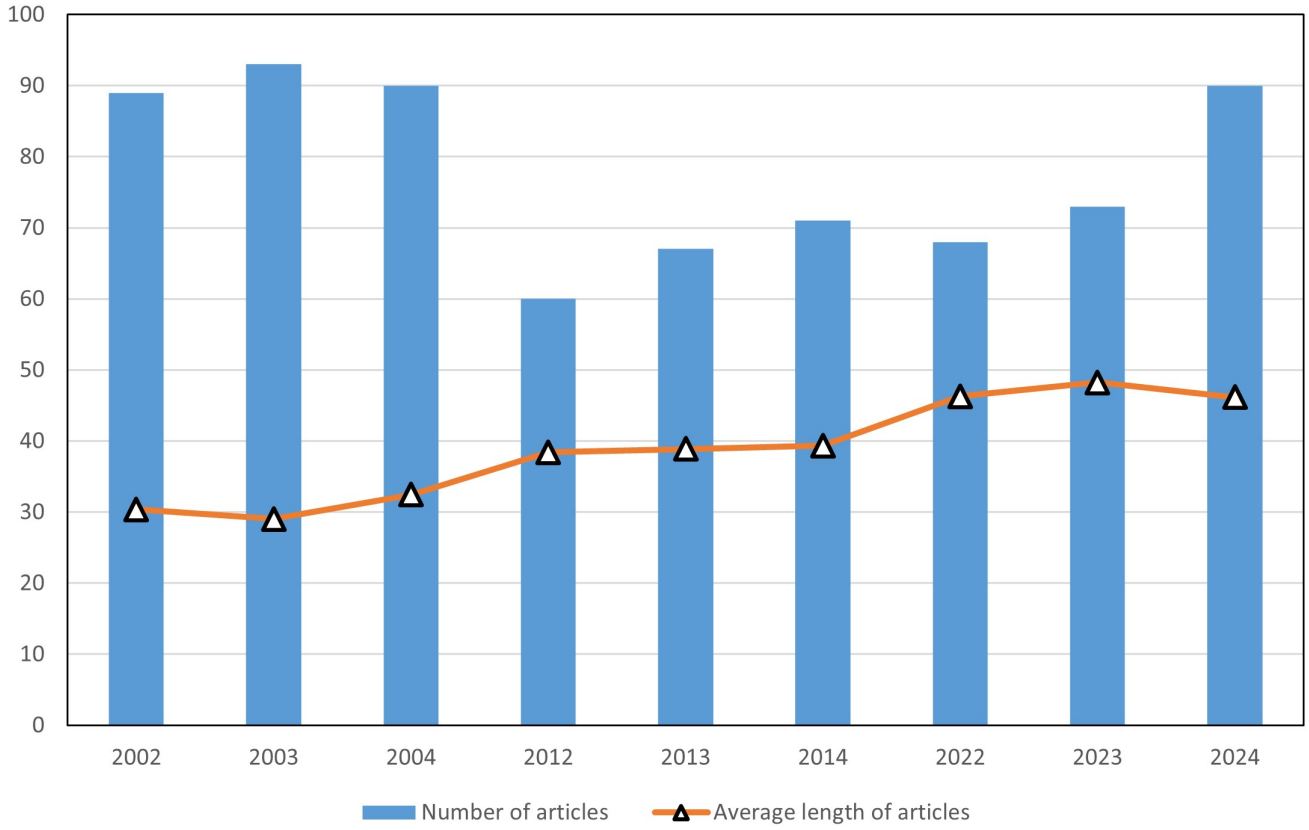

Figure 1. The number of research articles published and the average length of the articles (pages) each year.

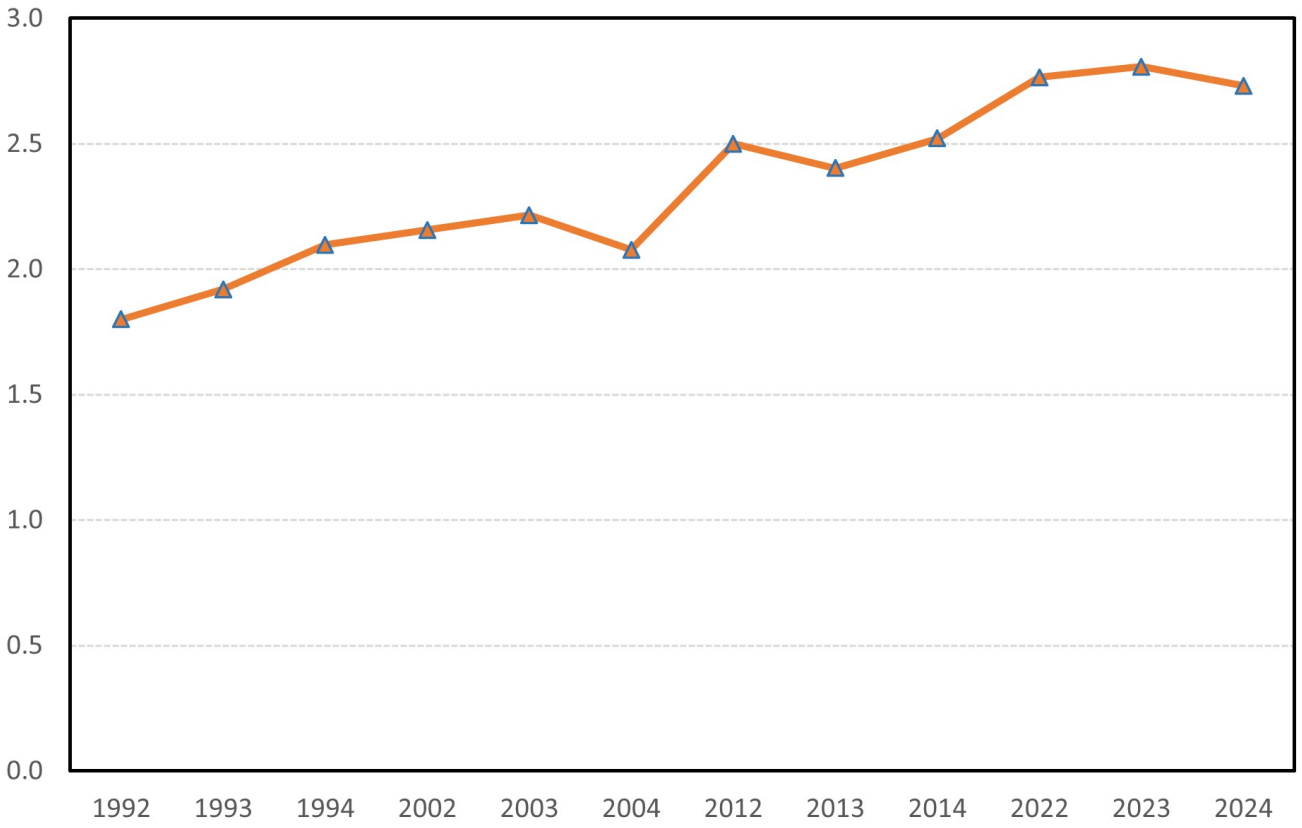

Figure 2. The average number of co-authors in research article each year.

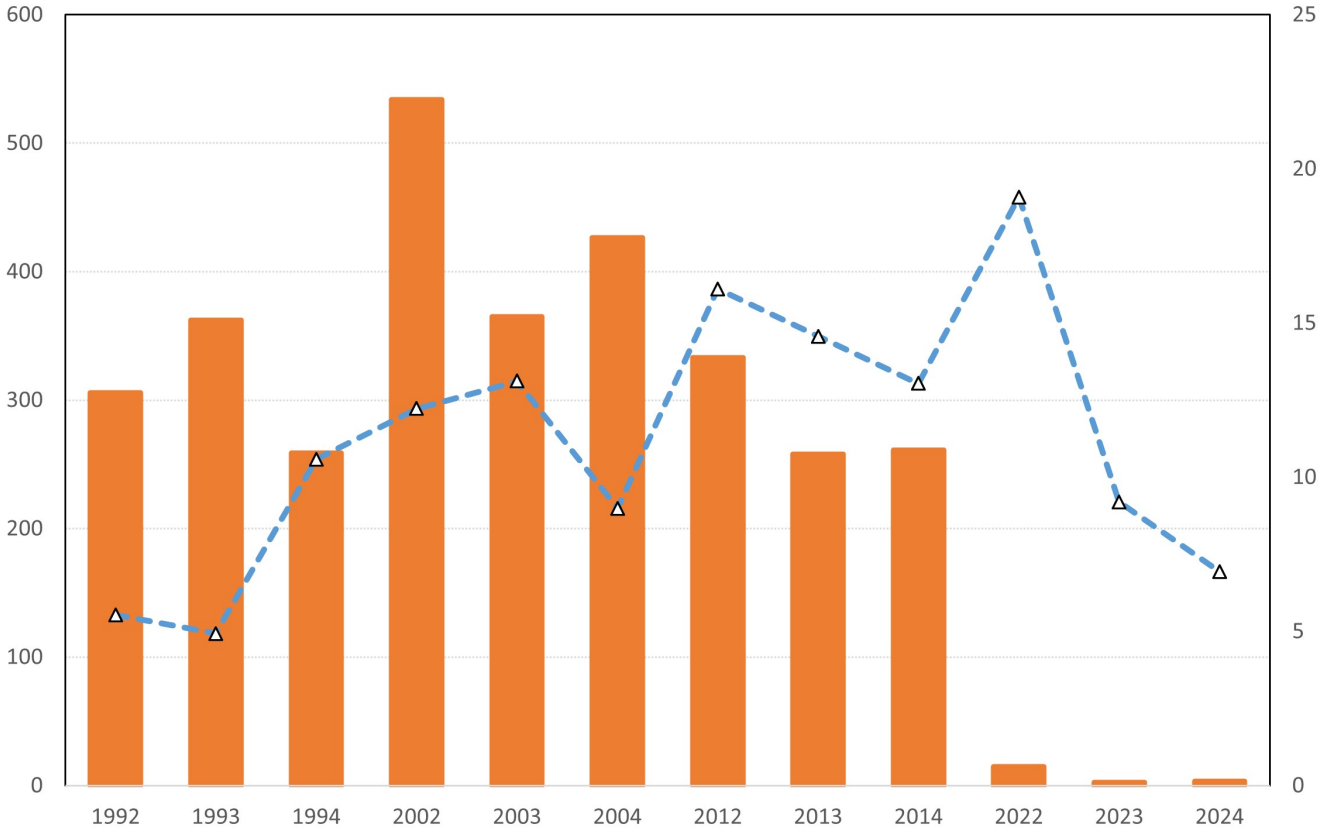

Figure 3. The average number of citations (bars, left y-axis) and altmetrics score (dashed line, right y-axis) for each research article.

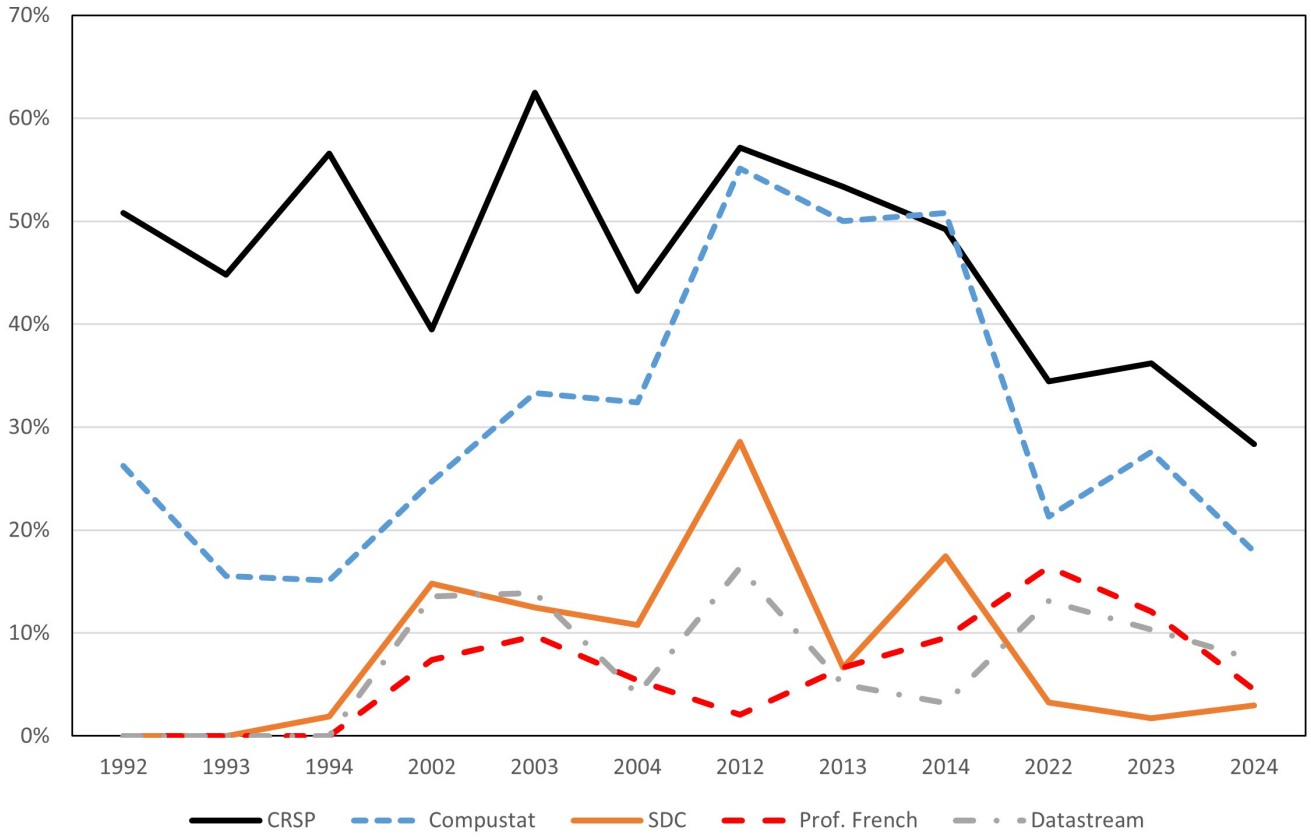

Figure 4. Popularity of the five most frequently used databases each year among articles that use empirical data.

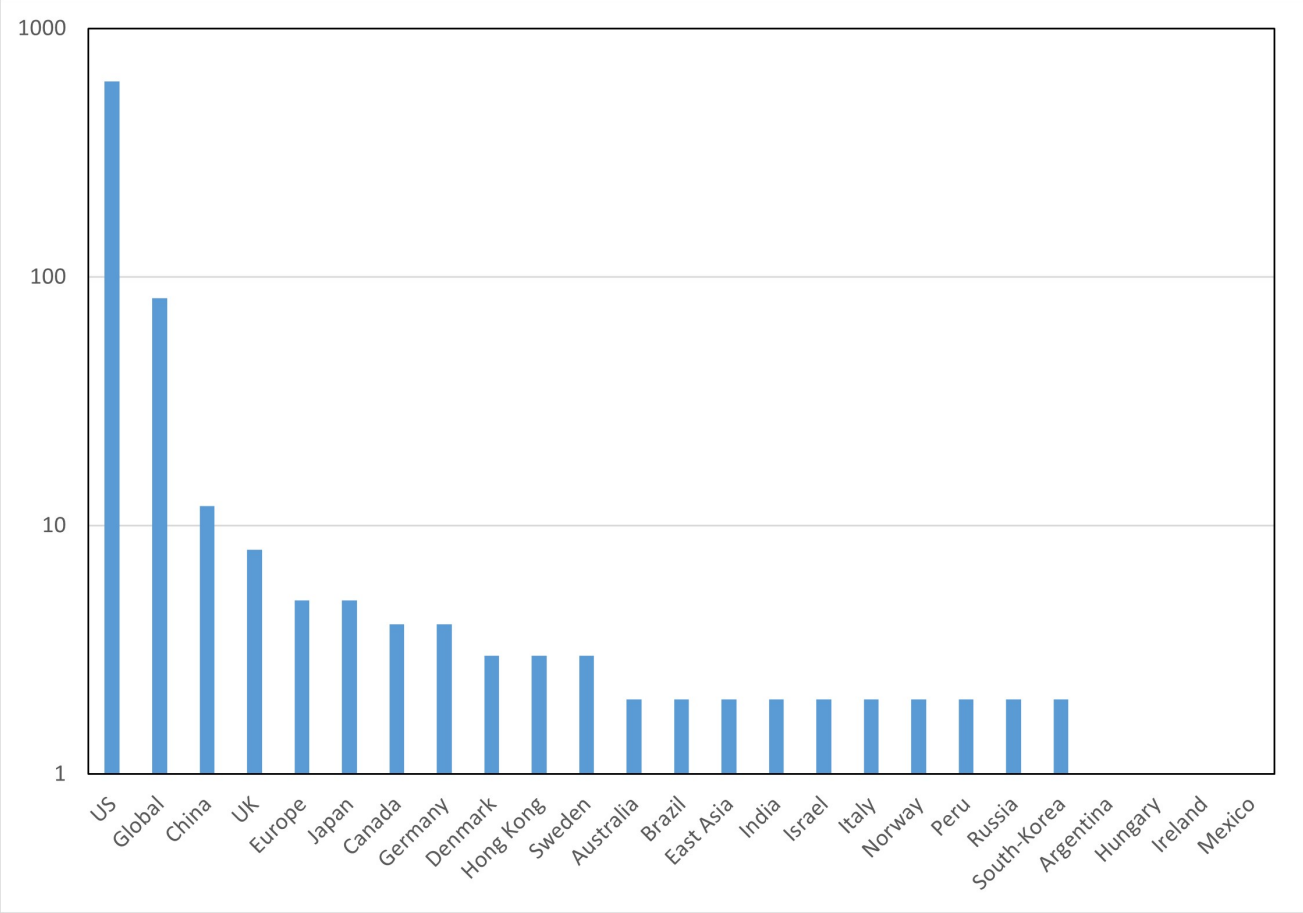

Figure 5. The number of times a given country or area is the main focus of the data used in the empirical articles. Note that the y-axis has a log scale.

Appendix A.

Table A - 1 Database descriptions. This table shows the databases that were most often encountered in the sample articles. The database descriptions are taken from the product's webpage. We have also estimated the annual cost for universities (USD).

Database name	Company	Main content	Price
CRSP	CRSP, LLC.	US stock market, risk-free rate, mutual fund data	$$$
CRSP provides the academic community, regulatory authorities, and investment practitioners with the preeminent financial data resources necessary to conduct innovative research, uncover scholarly insights, and inform sound investment decision-making.			
Compustat	S&P Global	USA & Global	$$$
Compustat is a database of financial, statistical, and market information on active and inactive global companies throughout the world. The service began in 1962. Compustat Financials provides standardized North American and global financial statements and market data for over 80,000 active and inactive publicly traded companies.			
Datastream	London Stock Exchange Group (LSEG)	World	$$$
Datastream is a global financial dataset that contains current and historical time series data on stocks, indices, bonds, funds, futures, options, interest rates, commodities, currencies, and economic indicators. It is accessible as part of several LSEG products, such as LSEG Workspace.			
SDC Platinum	LSEG	Corporate actions/World	$$$
Corporate deal information on everything from M&A to private equity to project finance. It is nowadays available via LSEG Workspace.			
Bloomberg	Bloomberg	Real-time data/World	$$$
Bloomberg terminal provides access to frequently used for financial market data, including e.g. stock prices, bonds, derivatives, and commodities.			
Prof. French	Professor French	Factor and portfolio data	free
The Data Library by Professor Kenneth French contains current benchmark returns and historical benchmark returns data, downloads, and details.			
I/B/E/S	LSEG	Analyst estimates/World	$$$
I/B/E/S database includes forecasts on earnings per share, revenue, cash flow, and other financial metrics, sourced from more than 19,000 analysts worldwide for over 40,000 public companies across 70 markets.			
EDGAR	U.S. SEC	Company filings/USA	free
The SEC's EDGAR database provides free public access to corporate information, most notably to companies' registration statements, prospectuses, and periodic reports.			
FRED	Fed	Economics series/USA	free
Federal Reserve Economic Data, FRED is an online database consisting of hundreds of thousands of economic data time series from scores of national, international, public, and private sources.			

Table I Descriptive statistics. This table shows descriptive statistics for various publication and authorship-related variables in the four subsamples. Note that all statistical estimates are averages of the articles in the three-year sample.

	1992-1994	**2002-2004**	**2012-2014**	**2022-2024**
Total number of articles	244	272	198	231
* Theoretical articles	72 (29.5%)	44 (16.2%)	26 (13.1%)	45 (19.5%)
* Open access / Free to read	9 (3.7%)	7 (2.6%)	8 (4.0%)	65 (28.1%)
* With Internet Appendix	0 (0.0%)	0 (0.0%)	100 (50.5%)	185 (80.1%)
* With replication code	0 (0.%)	0 (0.%)	0 (0.0%)	191 (82.7%)
Average number of pages / article	22.68	30.61	38.90	46.85
* Theoretical paper	21.97	28.80	39.65	49.31
* Empirical paper	22.98	30.98	38.79	46.26
* Shorter paper	14.76	19.00	*n/a*	*n/a*
* With Internet Appendix	*n/a*	*n/a*	39.13	46.45
Average number of authors / article	1.93	2.15	2.47	2.77
With 1 author	29.1%	22.1%	14.6%	12.6%
* 2 authors	50.0%	44.9%	35.9%	27.7%
* 3 authors	19.7%	29.0%	38.4%	34.6%
* $>$ 3 authors	1.2%	4.0%	11.1%	24.7%
Average number of citations	311.96	440.49	284.98	6.44
Minimum	2	9	9	0
Median	109.50	254.50	210.50	2.00
Maximum	6860	3507	1793	91
Average Altmetric Attention Score	6.81	11.46	14.48	11.23
Minimum	0	0	0	0
Median	0	6	10	3
Maximum	189	456	115	309

Table II Regression results for the citation counts. This table shows OLS regression results for the number of citations as a function of various variables. T-values are shown in parentheses below the parameter estimates. Huber-White-Hinkley (HC1) heteroskedasticity-consistent standard errors have been used in the estimation. Model 1 is a reduced-variable version of Model 2, based on Equation (2). Model 3 is similar to Model 2, but it is estimated as a Poisson Count model. Model 4 is similar to Model 2, but the dependent variable is $ln(1 + No\_cites_i)$. *** (**, *) denotes significance at the 1% (5%, 10%) level (two-sided test). Sample size is 944.

Variable	**Model 1**	**Model 2**	**Model 3**	**Model 4**
Constant	−576.077***	−616.445***	−0.363	−1.711***
	(−4.339)	(−4.171)	(−0.727)	(−6.019)
Years since publication	51.931***	52.620**	0.436***	0.548***
	(10.340)	(9.983)	(11.762)	(36.431)
(Years since publication)^2	−0.860***	−0.871***	−0.008***	−0.012***
	(−4.920)	(−5.005)	(−10.098)	(−29.923)
Number of authors	17.030	25.520*	0.101**	0.127***
	(1.124)	(1.868)	(2.232)	(3.166)
Internet Appendix	−24.806	−15.684	0.441**	0.447***
	(−0.843)	(−0.462)	(2.134)	(3.693)
Open Access / Free to read	186.277***	182.037***	1.019***	0.357***
	(3.198)	(3.280)	(7.698)	(4.971)
Number of pages	8.428***	8.284***	0.025***	0.028***
	(3.327)	(3.541)	(4.753)	(6.698)
Shorter Paper	−236.985***	−220.228***	−0.908***	−0.585***
	(−4.078)	(−3.877)	(−4.552)	(−3.780)
No Empirical Data	−83.069***	−89.727**	−0.323**	−0.368***
	(−2.428)	(−2.371)	(−2.333)	(−3.690)
Lead article in an issue		185.265	0.383**	0.447***
		(1.621)	(1.970)	(3.263)
Presidential Address		150.268	0.333	0.087
		(0.395)	(0.888)	(0.264)
$Adj.R^2$	15.8%	16.7%	32.9%	71.9%

Table III Regression results for the Altmetric Attention Score. This table shows the results of the AAS on various variables. T-values are shown in parentheses below the parameter estimates. Huber-White-Hinkley (HC1) heteroskedasticity-consistent standard errors have been used in the estimation. Model 1 is based on Equation (2). Model 2 estimates the same model for $ln(1 + AAS_i)$ with two new explanatory variables. Model 3 adds textual-frequency variables. *** (**, *) denotes significance at the 1% (5%, 10%) level (two-sided test). Sample size is 944.

Variable	**Model 1**	**Model 2**	**Model 3**
Constant	−6.665	−0.419	−0.438
	(−1.497)	(−1.399)	(−1.455)
Years since publication	1.008***	0.162***	0.158***
	(4.002)	(10.480)	(10.533)
(Years since publication)^2	−0.024***	−0.004***	−0.004***
	(−3.553)	(−10.944)	(−10.000)
Number of authors	0.728	0.057	0.071
	(0.864)	(1.299)	(1.614)
Internet Appendix	0.617	0.287**	0.275**
	(0.349)	(2.567)	(2.434)
Open Access / Free to read	4.937***	0.501***	0.469***
	(3.002)	(6.565)	(5.790)
Number of pages	0.260***	0.023***	0.021***
	(3.341)	(5.169)	(4.200)
Shorter Paper	−0.163	−0.096	−0.123
	(−0.071)	(−0.646)	(−0.825)
No Empirical Data	−5.633***	−0.296***	−0.262***
	(−4.747)	(−3.252)	(−2.817)
Lead article in an issue		0.334**	0.294*
		(2.042)	(1.793)
Presidential Address		0.416	−0.571
		(1.472)	(−0.406)
$Adj.R^2$	2.9%	20.7%	25.1%

Table IV Data sources and database usage. This table shows the percentages of theoretical and empirical articles across the four sample periods, followed by the overall average. Next, the breakdown of how many different data sources were used in empirical articles. It then shows how frequently the nine most commonly used databases were cited in the empirical articles. Databases are defined in the Appendix. Note that the total can exceed 100 percent, as many articles utilize multiple databases. Finally, the table shows the percentage of empirical articles that use data focused mostly on the US, global markets, or other markets.

	1992-1994	**2002-2004**	**2012-2014**	**2022-2024**	**All (%)**
Number of articles	244	272	198	231	945
Theoretical articles	29.5%	16.2%	13.1%	19.5%	19.8%
Empirical articles	70.5%	83.8%	86.9%	80.5%	80.2%
1 data source	36.6%	27.3%	13.4%	21.5%	24.8%
2 data sources	29.1%	28.6%	19.2%	16.1%	23.5%
3 data sources	16.3%	21.1%	26.7%	11.8%	19.0%
4 data sources	7.0%	13.2%	11.0%	11.3%	10.8%
>4 data sources	11.0%	9.8%	29.7%	39.3%	21.9%
CRSP	50.6%	47.6%	52.9%	32.8%	46.0%
Compustat	19.2%	29.1%	51.7%	22.0%	30.5%
SDC	0.6%	12.3%	16.9%	2.7%	8.5%
Datastream	0.0%	9.7%	7.6%	10.2%	7.4%
French	0.0%	7.0%	6.4%	10.8%	6.3%
EDGAR	4.1%	4.8%	9.3%	6.5%	6.2%
I/B/E/S	1.7%	4.4%	9.3%	7.0%	5.7%
Bloomberg	0.0%	3.1%	7.0%	9.7%	5.0%
FRED	0.0%	0.9%	0.6%	7.5%	2.4%
US focus	90.2%	81.4%	75.1%	72.2%	79.7%
Global focus	5.2%	11.7%	19.7%	7.5%	11.0%
Other	4.6%	6.9%	5.2%	20.3%	9.3%