Commercial data in financial research*

Marc Berninger[a], Florian Kiesel[b], Jan Schnitzler[b]

[a] *Department of Business Administration, Economics and Law, Technische Universität Darmstadt, 64289 Darmstadt, Germany*

[b] *Grenoble Ecole de Management, 38000 Grenoble, France*

Abstract

Databases are important but expensive sources for financial research. We examine data sources of 14,087 articles in 16 leading finance journals and identify the most common databases in empirical research. We find that 74% of empirical papers rely on commercial databases, combining on average three databases. In contrast to business users, who heavily rely on Bloomberg, CRSP and Compustat are the most relevant databases in financial research, used in approximately 30% of empirical papers, but this number doubles for top-5 journals. Articles using main databases on average receive 1.2 citations more per year, suggesting that data access enhances research clusters.

JEL classification: C8, D83, G00, G10

Keywords: Academic publishing; Commercial data; Databases; Data providers; Finance research; Citations

* Corresponding author: Florian Kiesel, Grenoble Ecole de Management, 12, rue Pierre Sémard, 38000 Grenoble, France; Phone +33 4 76 70 61 56; E-Mail: florian.kiesel@grenoble-em.com.

The authors are grateful to Patrick Augustin, Rui Dai, Campbell Harvey, Wei Jiang and Andrew Karolyi for helpful comments and suggestions on earlier drafts of this paper. All remaining errors are our own. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

1. Introduction

Having access to databases is nowadays fundamental for each finance and economics department. Modern computers allow academics to conduct empirical research by analyzing big data without much coding knowledge. Prior literature confirms the increased relevance of empirical research, with pure theoretical papers being less frequently published than in the past: Angrist et al. (2020) show that the use of data in academic economic research becomes more relevant as citations of empirical work increase as does the share of empirical papers. Hamermesh (2013) finds similar results by analyzing methodologies of articles published in three general economics journals between 1960 and 2010.[1] He likewise shows that the proportion of theoretical work has declined. Schwert (2021) examines the methodologies employed by papers published in the Journal of Financial Economics, one of the three leading finance journals. He shows that while in the first five years of the journal almost 60% of papers were theoretical, nearly 90% of the papers published between 2010 and 2020 have at least a small empirical part.

These trends indicate that financial researchers are increasingly required to have access to financial and accounting data. In contrast to most economic disciplines, where a lot of data is provided without further charges by national central banks, bureaus of statistics and other government or supranational agencies, access to financial data is dominated by commercial providers and therefore costly. Specialized firms, such as Refinitiv (formerly Thomson Reuters), Bloomberg or S&P Global, compile vast amounts of information, for which professional investors are willing to pay handsomely. Their services offer a wide range of different market data, such as

[1] Hamermesh (2013) examines 748 articles published in American Economic Review, Journal of Political Economy, and Quarterly Journal of Economics.

stock and accounting data, information on management board structure, financial news and even proprietary metrices based on their own calculations.

Even though there are usually discounts offered for academic purposes, the subscriptions to this data can be nevertheless very expensive.[2] In addition, the negotiations with commercial data providers can be complex and time-consuming. These challenges in obtaining relevant data for financial research is particularly problematic for small finance departments as the fixed costs for accessing data are high and the first user access being the most expensive one, while additional users can be added for lower rates or even for free. Therefore, synergies by sharing databases and their costs are hard to reach for smaller finance departments. Yet, also in bigger finance departments budget restrictions may lead to a reduced number of database subscriptions. This raises important considerations between the need to cut costs but still being able to provide proper conditions for high quality research. While the pricing side is usually determined in the negotiation of data acquisitions, its potential implications for the research output remain uncertain, also due to the lack of sufficient data and analysis, which makes cost arguments weigh in heavily.

In this paper, we present a ranking of the most common databases by investigating which data sources are mainly used by the academic community for financial research. Using the full text of more than 14,000 published finance articles, we identify the most common databases, describe for which research topics and countries they provide data and analyze how their usage impacts research and its visibility. Using modern machine-learning topic modeling, such as latent Dirichlet allocation (LDA) and other textual analysis methods, we divide articles into 14 different finance

[2] Subscriptions are often based on long-term contracts and lead to monthly costs which can easily exceed several thousand US dollars.

disciplines, proxy for the likelihood that a paper features empirical contributions, and whether it has a US or international focus.

We report several results: First, we support prior findings that the share of empirical contributions is growing, from about 70% in 2000 to almost 90% in 2016, stressing the importance of having access to the main financial databases. We create a list of 87 commonly used databases and show that 74% of all empirical papers in finance rely on at least one of them. The two most relevant databases are CRSP and Compustat as they were used in approximately 30% of all empirical finance papers that are published between 2000 and 2016. Interestingly, we find that in one out of two empirical papers published in a top five journal CRSP data is used, but only in 23% of the empirical papers in the remaining eleven journals. We observe similar numbers for Compustat data, indicating that the database used may serve as a proxy for the paper's quality or the potential to be published in leading finance journals.

This is supported by our finding that top business schools and US schools rely more on the Wharton Research Data Services (WRDS) platform than their peers, offering Compustat and CRSP data in an easy-to-download format. While the costs for databases are not publicly disclosed and depend on individual subscription packages, CRSP and Compustat are among the most expensive databases for academics. Our results indicate that cheaper alternatives, even though covering the required data as well, seem to be less convenient for academics and therefore less used by schools who have access to WRDS. This might be a result of data providers targeting primarily practitioners, leading to a competition between academics and business professionals in data needs as the latter ones need rather real-time information, but academics historical data.

Second, we find that it is often not sufficient to rely exclusively on one database. Most papers combine several data sources, and that number is increasing over time. As of 2016, we estimate that empirical papers use on average more than three databases. The number of databases also increases with the journal impact factor, suggesting that broad data access is a determinant for publication success. In addition, the number of robustness checks has increased significantly over our sample period, which is also associated with the number of databases used.

Third, we identify several clusters how multiple databases are frequently combined. CRSP and Compustat are the royal couple, often in combination with one or several additional databases. Governance data from various US regulations are particularly popular add-ons, indicating that these papers focus on the US market. We further find a second cluster which relies on international data by combing Datastream, WDI, Worldscope and Bankscope, yet the frequency of this cluster occurring is much lower.

Fourth, we examine whether a paper's visibility, measured by the average number of citations a paper receives per year, is associated with the database(s) used. Our results show that the visibility depends on the commonalty of underlying data sources. Academic articles using one of the 20 most common databases in our list receive on average 1.2 citations per year more than comparable articles. This effect remains robust, even after controlling for a wide set of control variables, including topic, year and journal fixed effects.

Our insights contribute to the literature in multiple ways: Kim, Morse and Zingales (2009) find that elite universities are losing their competitive edge (as physical access to productive research colleagues becomes less important). In addition, Agrawal and Goldfarb (2008) show that in particular middle-tier universities benefit due to easier communication channels and the possibility

of collaboration. However, for the finance discipline, Karolyi (2011) shows that the ratio of authors affiliated with top business schools remains stable over time in top finance journals. We find that top business schools and US institutions rely more frequently on expensive databases leading to a competitive advantage. By knowing and choosing the most relevant databases for their needs and without subscribing to not required data packages, smaller finance departments can specialize and contribute to outstanding research. Alternatively, researchers may gain access to data by co-authoring with scholars from larger and better funded institutions.

This paper furthermore provides a ranking of the most common databases and their usage in the last 20 years. By splitting our sample into 14 finance subdisciplines, researchers can easily assess which database is frequently used in their respective area. Our paper therefore also contributes to the strand of literature that is dedicated on publication outcomes and publication quality. While Kerl, Miersch and Walter (2018) and Reinartz and Urban (2017) evaluate the quality of academic finance conferences and rank conferences according to their appearance rates in top finance journals, we offer a comparable list of commercial databases for financial research and provide guidance on the quality and frequency using a first descriptive attempt and empirical evidence from more than 14,000 finance articles.

Two recent studies on empirical finance research are related to ours. On the one hand, Karolyi (2016) reveals a strong US-centric tilt in research: Only 16% of all empirical studies published in the top four finance journals use non-US data. One possible explanation for this result is that data availability and quality may not be evenly distributed across financial markets and countries. Karolyi (2016) mentions this point but does not analyze whether the bias towards US data is related to data availability. On the other hand, Dai, Donohue, Drechlser and Jiang (2021) examine 52,497

finance articles on Social Science Research Network (SSRN) and analyze several paper characteristics and their impact on publication visibility, measured by citations, downloads and journal outlet. They provide evidence that besides the paper's novelty the number of databases used in an article leads to more citations, downloads, and increases the chances to be published in a top three finance journal. Moreover, they show that the number of databases used in finance papers is increasing over time. While our paper confirms the increase in the average number of databases used and higher success chances for top publications, we also complement their study by showing that some databases are more frequently used than others. In particular, our results show that the databases used the most are also the most expensive ones.

We finally contribute to the question whether "more data" or "better data" is beneficial to enhance the visibility of research (Dominitz and Manski 2017). We find that using data from one of the most common databases is associated with more citations per year. This indicates that scholars tend to cite research that is conducted with common data. Our findings are in line with the *Union Heuristic* hypothesis. Harvey and Hirshleifer (2020) argue that authors are required to incorporate all referee suggestions to be able to publish papers and that this may prevent innovative research from being published. We contribute by showing that using the most common databases has a bigger success in conveying the paper's message, supporting the *Union Heuristic* hypothesis.

The remainder of this paper is structured as follows. Section 2 provides a brief background on commercial databases, their clients and an overview of common data providers and databases. Section 3 presents the data set and our methodology. The descriptive analyses of databases in academic journals are provided in Section 4, while Section 5 examines the impact of access to financial data on the research outcome. Section 6 concludes the paper.

2. Background on commercial databases

2.1 Database clients and changes in the database landscape

Data providers target different clients and academics are only one group interested in their data. The main customers of large data providers, such as Bloomberg and Refinitiv, are business professionals, e.g., investment bankers, analysts, and financial experts. While academics require stable and historical data for possibly all available firms in an easy-to-download format, business professionals usually need data for a comprehensive overview of up-to-date information. The latter preference can result in data discrepancies due to rewriting data which may improve the current information but leads to quality issues in the historical data. This may also lead to different conclusions depending on when the dataset has been downloaded. For instance, Ljungqvist, Malloy and Marston (2009) analyze the effect a change in historical data has on previous research outputs. The provider I/B/E/S offers data on historical analyst recommendations but implemented large-scale and nonrandom revisions. Ljungqvist, Malloy and Marston (2009) show that previous findings on return predictability stems to some extent from the problems of this I/B/E/S data. More recently, Berg, Fabisik and Sautner (2021) observe that Refinitiv ESG (formerly ASSET 4) rewrote their historical environmental, governance and social (ESG) ratings for firms and they provide evidence that this change in data has likewise an impact on the predictability of firms' stock returns.

In addition, the quest for new data leads to the creation of new databases and a competition between data providers, but also data providers change their services to adopt to the needs of their clients. As researchers are interested in analyzing new and reliable data, the current state of data can change. Subscriptions must be renewed, prices have to be renegotiated and scholars have to be aware of the most current database. One example of new financial data are spreads of credit

default swaps (CDS). CDS are a relatively new financial instrument traded over the counter, therefore data is hard to obtain, and researchers have to rely on data providers gathering the data directly from institutions. Mayordomo, Peña and Schwartz (2014) analyze CDS data offered by five commercial data providers and compare their corporate CDS prices. They show that there are differences among the data provided which makes it difficult for academics to decide which database gives the CDS market prices' most reliable account. The aim of this paper is to show which are the most common databases used in financial research since 2000, providing individual scholars but also academic institutions a general overview which databases are used by their peers and which databases are used for leading articles.

2.2 Overview of common financial databases

Financial research has early on developed a strong interest in empirical work, and extensive data analysis is now ubiquitous within the field. In contrast to large parts of economics, where high quality data is produced by the government or affiliated entities, many variables of interest to the finance community are based on proprietary information, and their owners successfully manage to market their products, often with the help of financial data providers. Even if raw data is publicly accessible, for instance through regulatory filings, its often decentralized and unstructured nature makes immediate usage prohibitively cumbersome such that the support of data providers becomes indispensable.

Thus, to enable a productive environment for financial research, it initially requires a significant investment into the data infrastructure. The first-choice database managers have to make is which provider to rely on to best access basic stock market data and firms' accounting information. There are several commercial providers who offer such a one-stop-shop service. Most convenient from

a researcher's perspective are data repository products, such as WRDS, that are specifically tailored for the academic use and combine access to multiple databases gathered by several data providers. Trauner (2017), however, shows that most academic institutions nowadays prefer to acquire data directly from data providers and do not use services from general platforms to eliminate the additional layer of cost.[3]

We list the most common products in Panel A of Table 1, also roughly sketching their data coverage. According to a recent FT article[4], Bloomberg holds the highest market share of 33%, largely thanks to their popularity among business customers. Refinitiv comes in second at 21% through a similar platform named Eikon, which includes access to Datastream and Worldscope which are more frequently used in academic research. Estimates for Standard & Poor's Capital IQ or FactSet are significantly lower and at about 5%.

[Insert Table 1 around here]

Looking from a pure academic perspective, this ranking is likely going to change due to the different needs between academics and practitioners. While most business professionals seek for a comprehensive overview of up-to-date information, academic researchers prefer large histories that are easily downloadable. Therefore, we also include CRSP, which is mainly accessed via the WRDS platform, to our summary in Table 1, despite its narrow focus on US stock market data, because it is so widely used among researchers. In combination with Standard & Poor's Compustat data, which can also be downloaded in bulk via the WRDS platform, it is a powerful tool quickly

[3] On the WRDS website is displayed that 525 institutions are subscribed to WRDS in 2021. QS listed approximately 26,000 universities on its website, indicating that approximately 98% of all universities anyway do not have access to WRDS.

[4] "Refinitiv's data-race struggle highlights LSE challenge", published April, 2020. https://www.ft.com/content/5f979bad-9b9e-46b1-b599-60c220bb8ffd.

put in operation for anyone with access to it. Finally, we also add Morningstar to our list in Table 1. This product is more frequently used for investment research as it contains detailed information about funds and other investment products. Yet, it consists of a comprehensive database also containing stock market information at the company level, making it to some extent a comparable product.

Once having access to such a core financial product, it can be complemented with other specialized data sources without limits, depending on the research objective. There are several other databases used in academic research that feature different objects or contain more specialized pieces of information. Some of them are offered by the same providers already mentioned, others are sold by smaller competitors or available even for free. Without the ambition to be comprehensive, we compile a second list provided in Panel B of Table 1 that contains other databases frequently used. Panel B of Table 1 also states what content each database is commonly used for. The convenience of WRDS as a one-stop-shop service can be seen in Column 3 of Panel B as several databases can be directly accessed through this platform. Most of the databases, however, need additional subscriptions and are not included in the basic WRDS subscription, leading to extra costs for academics. When identifying various data sources of articles published in finance journals, we employ an even more comprehensive list. For a complete list we refer the reader to Table A.1 in the appendix.

While it would be interesting to enlarge our findings with a detailed cost analysis of commercial data products, it is very challenging to achieve a fair point of comparison. The pricing list for research purposes is not transparent and often subject to negotiations between users and data vendors. Even if reliable prices were available for similar data products, there are still differences in the number of user accounts a subscription offers, how data is pre-processed and cleaned by a

provider, whether the service allows bulk downloading, and what data coverage a basic subscription entails. In fact, most vendors provide optional add-on services with more comprehensive data for additional fees. Integrating all of these contractual features into a detailed cost analysis is therefore more than challenging.

Yet, to provide at least a glimpse about the cost structure of financial data, we tried to collect subscription fees for basic access to the core financial products provided by the main data providers. Popular among practitioners are Bloomberg and Refinitiv's data terminals priced at about 20,000 USD per year. FactSet is available for around 12,000 USD. At an even higher end is access to CRSP and Compustat via WRDS, which is popular among financial researchers. Due to the issues mentioned before, this comparison should not be interpreted as whether a given data service is relatively cheap or expensive, a perspective unfortunately too often taken by librarians and managers in charge of data subscriptions. We rather want to remark that data access requires a sizable, fixed cost that creates problems to the research budgets of particularly smaller finance departments causing barriers to financial research.

3. Data and methodology

3.1 Sample

Our sample is based on the sample of Berninger, Kiesel, Schiereck and Gaar (2021). The sample includes 14,087 articles from 16 different finance journals published between 2000 and 2016 and contains the meta information of each paper, such as the authors' names, their affiliations, the paper's title, year of publication and its DOI. The advantage of the sample is that it additionally contains the full text of the articles which are obtained from the papers' portable document format (PDF) version. This allows us to search for the databases mentioned in the paper using a keyword search approach (detailed explanations of the keyword search is provided in Section 3.2). The

sample includes articles from the following 16 journals: European Financial Management, Finance & Stochastics, Financial Management, Journal of Banking and Finance, Journal of Corporate Finance, Journal of Empirical Finance, Journal of Finance, Journal of Financial and Quantitative Analysis, Journal of Financial Economics, Journal of Financial Intermediation, Journal of Financial Markets, Journal of International Money and Finance, Journal of Money Credit and Banking, Mathematical Finance, Review of Finance, and Review of Financial Studies[5]. The journal sample consists of general finance journals, such as Journal of Banking and Finance or Journal of Corporate Finance, but also contains journals that are rather specialized. While we include the top five journals, namely Journal of Finance, Journal of Financial Economics, Review of Financial Studies, the Journal of Financial and Quantitative Analysis, and the Review of Finance, we do not solely rely on the results of these five journals but also incorporate other journals to provide a more general view on academic publishing in finance but also to compare the findings between the top five journals and other journals. We made sure that all journals in the sample have a decent scientific reputation and are leading finance journals respecting by finance academics.

To guarantee the quality of the journal, the sample is based on quantitative and qualitative criteria (see Berninger et al., 2021): First, the five-year journal impact factors are obtained for all finance journals during the investigation period between 2000 and 2016. Practitioner-oriented journals (e.g., Corporate Governance: An International Review) and multidisciplinary journals (e.g., Financial Analyst Journal) were eliminated from the list. In addition, to exclude variations in the journal quality over time, journals need to have an average five-year impact factor higher than 1.45 for the sample period. Bryce, Dowling and Lucey (2020) show that the perception of researchers regarding the journal quality is not in line with national ranking systems and therefore

[5] Articles published in the Journal of Money, Credit and Banking and Financial Management are only considered since 2007 and 2005, respectively, due to a change in the publisher.

it is not always helpful to rely solely on citation impact factors. Berninger et al. (2021) therefore additionally consider the two journal rankings published by Currie and Pandher (2011, 2020) who survey finance researchers who have recently published on journal quality. Journals need to be listed at least as one of the best 30 journals in both lists to be including in the sample. Bajo, Barbi and Hillier (2020), Holden (2017), Karolyi (2016) and Smith (2004) who also analyze the content and meta information of finance journals and therefore need a sample of leading finance journals, have a comparable journal selection to ours.

3.2 Methodology

In this paper, we are interested in the databases authors of published papers in finance-oriented journals used for their research. To determine which database has been used, we apply a keyword search in the article's full text PDF. In a first step, we checked several university library websites from research institutions around the world, determined which databases they are subscribed to and created a comprehensive list of data providers and databases. We control for different writings of the database name (e.g., 13F, 13-F) but also for the name of the provider (e.g., IBES and I/B/E/S). We double checked our results using empirical papers without any hits for databases and added the databases mentioned in these papers if the database was commonly used. We present in our results only the 30 most common databases which is in line with the threshold that the database most be mentioned at least in 1% of all empirical papers in our sample. The entire list of databases can be found in Table A.1 in the appendix.

As databases are only required for empirical analyses and not for theoretical contributions, we follow the approach of Berninger et al. (2021) and estimate the likelihood whether a paper is rather theoretical or empirical by another keyword search approach in the full text of the article. This method assumes that empirical papers contain more words that describe the empirical design, such

as “dataset”, “variables”, and “descriptive statistics” and theoretical papers refer rather to words that are used to describe theory, such as “assumptions”, “postulations” or “theory”. As papers often cannot be contributed fully to one of the two methodologies, our approach is using a normalized measure by calculating the ratio of empirical and theoretical words:

$$Methodology = \frac{TheoryWords - EmpiricalWords}{TheoryWords + EmpiricalWords} \quad (1)$$

where *TheoryWords* are all words that are related to theoretical contributions and *Empirical-Words* is the number of all words that are related to empirical papers. The result is a value ranging between -1 (pure empirical) and +1 (pure theoretical). A value close to 0 indicates that the paper includes a theoretical part and an empirical one. These papers are therefore neither fully empirical nor theoretical. In this category belong theoretical papers with a small empirical analysis or empirical papers with (larger) theoretical backgrounds. As we are mainly interested in empirical papers, we define an empirical paper if the methodology variable is lower than -0.15 or lower than -0.05 but contains at least more than one table.[6]

Moreover, the use of databases may depend on the subject of the paper. Asset pricing needs for example stock return data, while papers on corporate governance require rather data on board structure and/or CEO characteristics. Berninger et al. (2021) control for topic-fixed effects by running a LDA clustering algorithm for 20 clusters. LDA provides for each article probabilities how likely it is that the paper belongs to one of the given topics. Berninger et al. (2021) define the paper’s topic as the one with the highest share, but they do not classify the 20 subjects and rely solely on the 20 clusters. We base our subjects on their approach but classify each cluster and

[6] We control the accuracy of this approach by randomly selecting a sample of 100 articles and analyze whether this approach correctly predicts the papers’ methodology. In most of the cases (around 90%) this approach led to the correct methodology.

combine some clusters as they are relatively similar. The output of the clusters and the corresponding topics are provided in Figure 1. We define the following 14 finance categories by merging some clusters into one: Microstructure (cluster 1), corporate finance (cluster 2 and 8), portfolio management (cluster 3), option pricing (cluster 4), asset pricing (cluster 5), financial econometrics (cluster 6 and 16), international finance (cluster 7), macro finance (cluster 9 and 15), IPO and M&A (cluster 10), banking (cluster 11 and 20), theory (cluster 12 and 17), fixed income (cluster 13), equities (cluster 14) and corporate governance (cluster 18 and 20).

In a related study, Baker, Kumar and Pattnaik (2021) analyze the topics published in the 25 years since the foundation of the Journal of Corporate Finance. They also conduct a cluster analysis based on common references and name nine clusters in the Journal of Corporate Finance.[7]

[Insert Figure 1 around here]

We further split the dataset in research articles examining US data and papers that have an international focus by using a keyword search and following the approach of Karolyi (2016). He provides a comprehensive list of keywords to compile international papers. We use his list and search for the related keywords in the paper's title and the abstract.

To determine the visibility of a paper, we measure the number of citations a paper receives in average per year. Citations are a common criterion to evaluate the visibility of researchers and the quality of articles (Netter, Poulson and Kieser, 2018; Bajo et al. 2020). We follow the process of Berninger et al. (2021) and Chan, Chan, Tong and Zhang (2016) and normalize the citations by dividing the total number of citations by the number of years since the paper has been published

[7] They name the clusters „private equity/IPOs", "corporate risk/mitigation", "corporate governance/executive compensation", "corporate restructuring/M&As", "family firms/corporate governance/ownership structure", "capital structure/corporate finance", "dividend policy", "corporate diversification/firm valuation", and "innovation/corporate social responsibility". As we have a sample of 16 finance journals, our clusters are more general, but we cover all nine clusters.

as papers that are longer outstanding are likely to receive more total citations. We obtain the numbers of citations from Crossref and Google Scholar, both citations counts were downloaded in July 2020.

4. The use of databases in academic journals

4.1 Overview of data by journals

Our dataset includes publications of 16 academic finance journals. As shown in Panel A of Figure 2, the number of published articles has been steadily increased from about 400 in 2000 to more than 1,300 in 2013. This increase is largely explained by an increase in the number of articles accepted by journals, even though Financial Management and the Journal of Money, Credit, and Banking enter our data only in 2005 and 2007, respectively, which also contributes to this increase.

[Insert Figure 2 around here]

In the same figure we also plot the percentage of published empirical papers identified through our algorithm. The share of empirical work has likewise been growing over our sample period, starting at about 70% in 2000 up to almost 90% in 2016. These numbers compare well to the findings of Schwert (2021) for the Journal of Financial Economics, stating that 88% of papers contained an empirical part over the past decade.[8]

Using the subsample of all empirical papers, Panel B of Figure 2 plots time-series evidence about the database identification of our methodology. The percentage of papers where we identify at least one database from our list increases from 60% in 2000 to more than 80% in 2016. We have

[8] We consider 1,617 articles from the Journal of Financial Economics. This is approximately 11.5% of our total sample and therefore our results are not driven by this journal.

two explanations for such an increase: first, seeking conforming evidence or simply for convenience, researchers converge to employing the same data. Alternatively, increasing demands to deliver additional control variables and robustness tests, or just to get into a position to build on established evidence in existing literatures, requires researchers to compile more data sources.

The figure also reports the number of databases used. To have a clear-cut distinction from the previous graph, we calculate these statistics conditional on observing at least one database. The average number of databases increases from less than two to more than three during our sample period. Such an increase is suggestive of the latter explanation where researchers are required to compile more data. The effect becomes even more pronounced when we look at the top five finance journals only. There, the average number of databases used sharply increases from 2 to more than 3.5 towards the end of our sample.

Table 2 reports the percentage of published empirical papers for each journal separately. There is significant variation in the ratio of empirical contributions across the journals in the sample. In support of our algorithm, we find a strong correlation with a journal's reputation to support mathematical or theoretical work. For instance, the lowest share of empirical work is published in Finance and Stochastics (34%) and Mathematical Finance (37%). In contrast, Financial Management (98%), the Journal of Empirical Finance (97%) and the Journal of Corporate Finance (93%) mainly publish empirical papers or theoretical contributions with a large empirical part. The remaining journals show percentages between 75% and 90%.

[Insert Table 2 around here]

Table 2 also introduces first statistics about our main variable of interest. Column 5 ("% only emp.") indicates the percentage of empirical papers for which we are able to identify at least one database from our list of 87 databases. On average we find that 74% of all empirical papers use at

least one of these databases. These numbers range from 13% for Finance and Stochastics to 93% for Financial Management. While hardly covering an exhaustive list of databases, in particular since researchers aim to work with unique and interesting data sources that are hard to capture individually, these statistics show that researchers repeatedly rely at least partially on the same sources. If we look at the percentage among all published papers in Column 6 ("% all"), we obtain somewhat smaller numbers, yet the overall picture remains that the main commercial databases are widely used.

As already shown in Panel B of Figure 2, many of these databases are not used in isolation but in combination with each other. Column 7 ("Avg. #") reports the average number of databases for each journal conditional on that we are able to identify at least one database. For journals with an empirical focus, we observe on average more than three databases per paper. Column 8 ("Std. Dev.") reports the corresponding standard deviation.

4.2 Commonly used databases in financial research

In this section we take a closer look at which databases are most frequently used in published finance articles. At the top of the ranking of the most commonly used databases shown in Table 3 are CRSP and Compustat, which represent the default option to combine US stock returns with firms' accounting information for academic purposes. CRSP shows up in 34% of all empirical papers in our sample (see column "% empirical") and Compustat in 27.5% of the cases, respectively. These high numbers indicate that there is a strong tilt towards research about US markets in top finance journals, one reason being that it represents the largest stock market enabling a maximization of sample sizes. It also suggests that this combination captures a large market share as there are at least theoretically other databases with similar data available, yet the access may be less convenient than the one through the WRDS platform. Our results are supported by the findings

of Karoyli (2016) who shows that only a small part of empirical papers published in the top four finance journals examine non-US data.

[Insert Table 3 around here]

Datastream takes the third rank as it appears in 16% of all empirical papers. At least historically, it was the prime source for stock returns of international companies, yet it contains much more data that could help contribute to its common usage. Similar products with a broad coverage of pricing information for all types of securities, commodities and indices are Bloomberg ranked 5th (9.5%) and recently FactSet ranked 37th (0.6%). In addition, there is Capital IQ on rank 34 (0.9%), which is hard to rank as it also contains Compustat, even though most researchers might access Compustat via WRDS.

Well ranked in Table 3 are also specialized databases that managed to establish themselves as default options for specific pieces of information. Among them are SDC as a source for mergers, security issuance and other transactions (ranked 4th with 11%), I/B/E/S for stock analysts (ranked 6th with 7.4%), Execucomp for data on executives and their compensation (ranked 8th with 4.9%), ISS/RiskMetrics for governance data of directors, and shareholder voting and proposals (ranked 9th with 4.6%), and TAQ for intraday transactions on NYSE (ranked 10th with 4.6%).

There are also non-commercial data sources among the most used databases. Most notably, International Financial Statistics (IFS) provided by the IMF (ranked 7th with 6.2%) and SEC's EDGAR (ranked 11th with 4.4%), which contains unprocessed versions of regulatory filings for listed firms in the US. Finally, there are news search engines like Factiva (ranked 16th with 3.6%) and Lexis-Nexis (ranked 27th with 1.6%). Table A.1 in the appendix lists all remaining databases we searched for that did not make it into the top 30 most commonly used databases.

The remaining columns of Table 3 separately report the percentage for each database according to various sample splits. First, we divide the sample into whether one of the authors is affiliated with a top ten business school or not[9]. We find that on average papers associated with top business schools are more likely to make use of the common databases on our list of 87 databases. This seems to be reasonable as these school usually have access to all common databases. The effect is particularly pronounced for CRSP (46% vs 33%), Compustat (36% vs 26%), and CDA/Spectrum (6.4% vs 3.9%). In contrast, researchers from other schools are more likely to use Datastream (16% vs 10%), Bloomberg (9.6% vs 7.9%), IFS (6.5% vs 3.9%), and Bankscope (3% vs 0.7%), which might be cheaper solutions and cover firms around the world and are therefore not focused on the US market.

When splitting the sample into top five journals and other finance journals, very similar patterns arise even though the effects become more pronounced. For example, the difference between these two groups increases to 51% vs 23% for CRSP and 41% vs 19% for Compustat. When splitting the sample into award-winning papers and other publications, it yet again amplifies the effect. On the other hand, the pattern that Datastream, Bloomberg, IFS, and Bankscope are more likely to be used in each respective other group prevails.

Several explanations are potentially able to contribute to such a finding. Better data quality and more stringent disclosure requirements enable more impactful research about US markets, for which CRSP/Compustat are the prime data sources. Arguing from a negative point of view, biases within the academic community could favor US-centric research. Finally, high fixed costs for some

[9] We follow the definition of Berninger et al. (2021) for business schools using the Financial Times' Global MBA ranking in the year prior to the publication. An article is considered as from a top business school if at least one of the authors is affiliated with such an institution.

of the top items on the list makes data access prohibitively expensive for smaller institutions, forcing their researchers to switch to other data sources or topics. Related to the latter point, we also think that this paper produces valuable statistics that should enable more informed decisions about which databases are worth the investment.

Based on the 10 most common databases in our sample, we plot their time-series development in Figure 3. The first impression tells us that market shares are fairly stable where each database enjoys steady academic interest over time. CRSP ranks first for every single year with a modest increase from 30% to 35% of papers. Compustat started out at less than 20% in 2000, but it has been strongly catching up in most recent years. Most striking is the relative increase of Bloomberg's prevalence in academic research. Starting at less than 5% of papers in 2000, its occurrence has more than tripled to almost 20% in 2016. We also report strongly increasing percentages for SDC, IBES, and ISS/RiskMetrics in recent years.

[Insert Figure 3 around here]

Next, we break down our analysis into finance subdisciplines based on research topics. Following the word clouds introduced in the methodology section, we assign each paper to one of the 14 research topics. Then we plot in Table 4 for each topic how frequently each database is used. When selecting the databases included in Table 4 from our list, we make sure that for each financial topic at least the top five databases are provided. These more detailed statistics are of interest because most researchers are specialized on one or few areas where very different types of data may be needed. In addition, this analysis provides a good sanity check for our data quality since there might be strong priors about which database should be associated with which topic.

[Insert Table 4 around here]

The results suggest that we achieve a fairly good mapping between topics and databases. For example, Bankscope, which appears in less than 3% of all empirical papers, suddenly becomes the dominant data source at almost 20% if we only consider banking related papers. Similarly, if we look at the market microstructure field, high-frequency data from TAQ becomes very important at 34%, which otherwise only appears in 4.6% of all cases. Generally speaking, corporate finance-oriented topics, including governance and IPOs/M&As, heavily rely on Compustat, SDC, I/B/E/S, and Execucomp, whereas CRSP, Datastream, Bloomberg are more relevant for asset pricing fields.

Finally, Table 5 shows the correlations between the top 20 databases. The table reports the likelihood for each database to appear conditional on observing the one depicted in each respective row. The most evident observation in Table 5 are several frequent clusters consisting of three databases that combine CRSP and Compustat with one additional database. These are in particular SDC Platinum, I/B/E/S, Execucomp, ISS, EDGAR, CSA/Spectrum and Factiva. The likelihood of observing an additional fourth database is already significantly diminished. A second cluster of international data sources combines Datastream, WDI, Worldscope, and Bankscope, yet its prevalence is significantly reduced in comparison to the US clusters.

[Insert Table 5 around here]

4.3 The main data providers

In this section we slice our data into different observational units. Instead of looking at marketed databases, we aggregate it at the level of managing companies. Recently, there have been strong signs of industry consolidation, making financial data providers interesting targets for even larger financial services groups as prominently shown by London Stock Exchange Group's acquisition of Refinitiv in 2020. Thinking about market power, it is informative to see what percentage the largest players take in the academic niche of the market.

We collapse our data by assigning all databases to its current owners, irrespective of historical changes that might have taken place. Table 6 reports the percentage of publications relying on different financial data providers, as well as separate statistics for each subfield introduced. Refinitiv, through its various data offerings, manages to topple CRSP from the top position with 36% vs 34%. Standard & Poor's also stays within reach at 29%. The fourth largest data provider is Bloomberg with 9.5%, after which the numbers quickly start to fall.

Comparing the main player's coverage of different topics, Refinitiv appears to have a strong balance between all of them. CRSP, largely due to its flagship product of US stock returns, also manages to generate usage across various disciplines. Standard & Poor's coverage appears relatively stronger for topics related to corporate finance, whereas Bloomberg is particularly in research of options and fixed-income markets.

[Insert Table 6 around here]

5. Access to data and financial research outcome

5.1 Institutional resources and data choice

In this section we first attempt to provide suggestive evidence for our hypothesis that research, in particular data choice, is affected by funding constraints. We reduce our sample to publications using one of the core data packages introduced in Panel A of Table 1: FactSet, Refinitiv's Datastream/Worldscope, Bloomberg, Compustat or S&P CapitalIQ, and/or CRSP. The idea is that these sources are frequently used to collect firms' stock prices or accounting information, making them to some extent substitutes to retrieve similar types of data. As discussed in Section 2, the fixed cost to obtain institutional access differ significantly with FactSet being at the lower end and Compustat and CRSP at the upper end. Use of the latter data services are a strong indication that

the authors have access via WRDS, which comes at additional cost but offers data verification and a lot of convenience in terms of sample construction.

Based on the assumption that authors with a choice opt for the research-tailored services provided by WRDS, we estimate a multinomial logit regression aiming to explain the choice of data. The independent variables of interest are two dummy variables capturing information about the affiliated institutions of authors, top business schools and US schools, both used as proxies for the research funding available. Our regressions also control for the number of authors, as larger teams might jointly have more data choices, and a dummy indicating whether a paper studies an international or non-US sample. The latter is an important control variable because CRSP does not cover stock returns of international companies.[10]

The results are reported in Table 7. In comparison to papers with Bloomberg data which we selected as the benchmark group, papers using Compustat or particularly CRSP data are significantly more likely to be written by authors affiliated with top business schools or US institutions. The reported coefficients for top business schools translate into relative risk ratios that have a 1.5 times higher likelihood for Compustat data and a 1.6 times higher likelihood for CRSP data. The corresponding numbers for US schools are 3.0 and 3.4. Differences of FactSet and Refinitiv data with respect to the Bloomberg benchmark are less pronounced. The only significant coefficient is that authors with US affiliations are relatively less likely to publish work with Refinitiv data.

The final column of Table 7 presents an alternative specification from an ordered logit regression. In this model we create a variable that ranks the five core databases according to its estimated fixed cost (FactSet=1 and CRSP=5). Again, we find that authors affiliated with top business

[10] Dropping all publications with international focus gives very similar results, which is why we report the results for the more comprehensive sample including the international dummy as a control variable.

schools and US institutions are more likely to retrieve data from sources coming at a higher fixed cost. Importantly, we would like to stress that the presented results should not be interpreted as causal estimation of the effect since the underlying data is measured and there might be alternative explanations for it. Our goal here is rather to produce first stylized evidence suggesting that data choices in financial research are linked to the researchers' affiliation, which may potentially also affect research outcomes.[11]

[Insert Table 7 around here]

5.2 Number of databases and publication outcome

Next, we examine whether access to data, measured by the number of databases, has a potential impact on papers' publication outcomes. The ranking of a journal that eventually publishes a paper is of particular importance to researchers as their tenure-decisions, research time and publication bonuses frequently depend on it. Therefore, we test whether more databases are associated with publications in journals with higher impact factors. We have already shown in Figure 2 that papers published in top 5 journals use on average more databases. Here, we test more rigorously whether such a positive correlation remains in regression analysis after controlling for other factors.

Since the journal impact factor only varies across journal-year observations, we decided to use the number of databases as the dependent variable in our regressions. After adding a large set of predictors to the regression model to fit the variation in the number of databases, does the journal impact factor still have predictive power? Using the sample of empirical papers, we report the regression results in Table 8.

[11] Relatedly, we also estimate whether research topic choices are similarly affected. The results are difficult to interpret as they are subject to even more confounding explanations, but they could indicate that resources and data access are associated with research topics. We report them in Table A.2 in the appendix.

Model 1 reports the bivariate correlation between the two variables of interest, which is positive and highly statistically significant. Controlling for year and topic fixed effects reduces the coefficient from 0.44 to 0.36, but it remains statistically significant. A one-point increase in the impact factor, which for instance corresponds to the difference between the Journal of Empirical Finance and the Review of Finance in the most recent data, is associated with 0.36 more databases on average.

In Model 3 we add three additional variables of interest to the regression: first, a dummy variable indicating whether a paper features robustness tests or an entire robustness section.[12] We expect that this should trigger additional data needs and thus predict a positive coefficient. This is indeed the case and the coefficient is statistically significant in all specifications. Using the smallest point estimate, we find that on average every fifth paper with robustness tests requires an additional database. Second, we include the log of each paper's number of citations per year to control for the quality and impact of each paper. This is a controversial control variable as it also has a mechanical effect on the journal's impact factor in which a paper is published. Yet, the correlation coefficient between the two variables is 0.49 and the inclusion of this variable yields more conservative estimates. Either way, the journal impact factor continues to have a significant relation with the number of databases used. Third, we include a dummy variable indicating whether the paper uses international or non-US data. This coefficient has a strong negative and statistically significant effect on the number of databases employed, indicating that there are more data sources available for the US market which can be linked with each other. The results extend the findings

[12] In order to analyze whether a paper contains robustness tests, we ran a keyword search in the main body of the document, using related keywords, such as "robustness check", "further tests" or "robustness section". The approach is similar to Berninger et al. (2021).

of Karoyli (2016) who finds that research in international finance might be underrepresented due to the uneven quality of data in non-US markets.

We add the entire set of additional control variables from Berninger et al (2021) in Model 4, which reduces the effect but does not eliminate it. In Model 5 we exclude publications in Mathematical Finance and Finance & Stochastics as they infrequently publish empirical work, yet again the results are not driven by this. Thus, we conclude that there is a robust link between the number of databases employed in research and the publication success in terms of the journal impact factor.

Related to our discussion here is the analysis in Dai et al (2021). Using a large sample of working papers circulating on SSRN, they estimate determinants explaining their chances of getting published in a top three finance journal. They also find that the number of databases has a positive effect. In addition, they investigate the number of citations as a dependent variable, which is what we will do in the next section.

[Insert Table 8 around here]

5.3 Databases and their impact on citations

We finally examine whether databases used in a publication affect the visibility a paper generates. Several papers examine the impact of article characteristics on the article's impact, proxied by a publication's average number of citations per year. Karolyi (2016) finds that papers which examine non-US data receive on average more citations, while Dai et al. (2021) show that the number of databases used has a significant effect on the number of citations. Ex ante it is not obvious whether we should expect a positive or a negative effect for the most used databases. On the one hand, if the underlying data is widely accessible, the research can be easily replicated, putting it under more scrutiny and leading to more citations. Alternatively, one could assume that the most common databases distributed by commercial vendors have a negative impact as there is

more potential for an innovative contribution with the help of unique, proprietary data. The results contribute to our understanding whether standardized databases have a positive or negative impact on the visibility of articles and consequently extend the findings of Karolyi (2016) and Dai et al. (2021).

We employ regression analysis where the dependent variable is the number of citations per year. We construct two variables of interest, one indicator whether one of the main 20 data sources was used, and one indicator for any of the other data sources from our list. Thus, our control group consists of publications where no data source has been identified (despite it was flagged to be an empirical paper). We attempt to control for other effects influencing the quality of a given publication. In doing so, we closely follow Berninger et al. (2021) by including author-specific variables, such as top scholars and US or top business school affiliations, as well as article-related variables, such as lead articles, award winners, and special issues. These control variables are explained in more detail in the table. We also include fixed effects for publication years, financial topics, and journals. The results are shown in Table 9.

[Insert Table 9 around here]

Without any controls, we find that papers in which we identify data sources generate 1.7 citations more per year using the Crossref citation database. If any of the main databases is used, this difference is even bigger at 3.5 citations per year. After including our full set of control variables in Model 4, the differential effect from less frequently used databases is close to zero and becomes insignificant. Publications using one of the main databases, however, continue to generate 1.2 citations more per year and this effect is also statistically significant. Using Google citation in Columns 5-8, the observed differences are even more pronounced. Our specification with the full set

of control variables estimates that papers using the main data sources generate 2 citations more per year.

Two additional variables of potential interest are the dummy variables indicating the existence of robustness sections and the use of international data. Having robustness sections is associated with 1-2 additional citations per year, depending on the specification. Interestingly, international data is associated with relatively more citations once we control for the full set of control variables.

In Table A.3 in the appendix we report additional results where we separate the effect for the main databases into their individual components. While on average positive, the point estimates for individual databases vary. For instance, SDC and Morningstar tend to be associated with fewer citations than our control group. In contrast, Compustat, Worldscope, CBOE VIX, and Bankscope receive relatively more citations. In particular the latter may be evidence for overly coarse topic classifications, yet the descriptive statistics may nevertheless be interesting for some researchers working specifically with this data.

While there is an own research strand focusing on the impact of research of scholarly literature - "sciencometrics" - and how to measure its impact, there are only few papers that focus on the finance discipline in particular. Ederington (1979) finds that researchers from elite business schools receive approximately 70% more citations than researchers from unranked schools. Klemkosky and Tuttle (1977) show that only six university were responsibly of approximately 25% of the total pages in leading finance journals until 1975. Heck, Cooley and Hubbard (1986) find that 201 institutions are represented in the Journal of Finance between 1966-1975, but the number of institutions increased by 34% to 270 institutions between 1975 and 1985. We contribute to the research agenda on determinants of citations and the success determinants of academic publishing

by showing that relying on the most common databases increases the visibility, measured by citations per year, of publications, extending the findings of Dai et al. (2021) who show that the average number of databases leads to more citations. Moreover, our findings are in line with the *Union Heuristic* hypothesis as we find that only the mostly common databases have a positive impact on the number of citations a paper receives, suggesting that referees, editors and readers might be more convinced by results that are gathered with established data sources. Harvey and Hirshleifer (2020) argue that editors rely on the *Union Heuristic* and that papers need to address all requested tests and extensions to be published.

6. Conclusion

Databases are a crucial factor for empirical finance research. Commercial databases provide different data in an easy-to-download format, but this data can be very expensive. These costs might be hard to handle for smaller finance departments as they do not benefit from economies of scale, but also cost-efficiency gains may require less database subscriptions for larger departments. Moreover, academic institutions are not the main target group for most data providers. Data providers adopt to the needs of business professionals, which may have consequence for the academic use and changes in the academic use of these platforms.

In this paper, we provide a ranking of the most common databases in financial research by assessing the frequency with which they have been used in published academic articles between 2000-2016. We find that 74% of all empirical papers use at least one of these databases, but the numbers vary depending on the journal. While CRSP and Compustat are still the most used databases in financial research, used in 34% and 27.5% of all empirical papers, respectively, we find a relative increase of Bloomberg's prevalence in academic research. Our results also reveal that databases are used in combination, and we find two main clusters. The first cluster focuses on the

US market by combining CRSP and Compustat with one additional database, while the second cluster is targeting international samples by examining data obtained from Datastream, WDI, Worldscope and Bankscope, respectively. We also find that the data used highly depend on the research subject and the affiliated institution. Bankscope, for example, which appears only in 3% of all empirical papers, is the dominant data source for research in banking. Similar effects can be found for high-frequency data from TAQ, which appears in less than 5% of all empirical papers but is one of the most used databases for research in microstructure. The results also indicate that researchers from top business schools and US schools have 1.5 times and 3.0 times higher likelihood for using Compustat and CRISP data, respectively, indicating that authors affiliated to these institutions are more likely to retrieve data associated with higher fixed costs.

Further results reveal that the use of databases affects the visibility of a paper, measured by the number of citations per year. We find that papers relying on the 20 most common databases receive on average 1.2 citations per year more than a similar article. The result is robust using year, journal and topic fixed effects and to the source of citation numbers. We, however, find that already the next best databases have limited impact on the number of citations, indicating that only the main databases play a significant role in the visibility. These results extend the findings of Dai et al. (2010) who show that, among other criteria, the number of databases used increases the article's visibility. Our results are in line the *Union Heuristic* hypothesis as we find that using the most common databases have a bigger success in confirming the paper's message rather than showing the same results with new data.

This paper provides a first descriptive attempt examining empirical data of 16 finance journals and more than 14,000 finance articles to provide a systematic overview of commercial data in financial research. While these results might be helpful to assess which subscriptions are beneficial

for researchers, they do not attempt to identify a causal effect between databases and article quality. We solely argue that common access to data enhances the visibility of publications. Thereby, we focus exclusively on the most common databases and forego to analyze which other data has been used in published studies.

References

Agrawal, A., & Goldfarb, A. (2008). Restructuring research: Communication costs and the democratization of university innovation. *American Economic Review*, 98(4), 1578–1590. doi: 10.1257/aer.98.4.1578.

Angrist, J., Azoulay, P., Ellison, G., Hill, R., & Lu, S. F. (2020). Inside job or deep impact? Extramural citations and the influence of economic scholarship. *Journal of Economic Literature*, 58(1), 3–52. doi: 10.1257/jel.20181508.

Bajo, E., Barbi, M., & Hillier, D. (2020). Where should I publish to get promoted? A finance journal ranking based on business school promotions. *Journal of Banking & Finance* 114, 105780. doi: 10.1016/j.jbankfin.2020.105780.

Baker, H. K., Kumar, S., & Pattnaik, D. (2021). Twenty-five years of the journal of corporate finance: a scientometric analysis. *Journal of Corporate Finance*, 66, 101572. doi: 10.1016/j.jcorpfin.2020.101572.

Berg, F., Fabisik, K., & Sautner, Z. (2021). Rewriting history II: The (un) predictable past of ESG ratings. *European Corporate Governance Institute Working Paper*. doi: 10.2139/ssrn.3722087.

Berninger, M., Kiesel, F., Schiereck, D., & Gaar, E. (2021). Citations and the readers' information-extracting costs of finance articles. *Journal of Banking & Finance*, 131, 106188. doi: 10.1016/j.jbankfin.2021.106188.

Bryce, C., Dowling, M., & Lucey, B. (2020). The journal quality perception gap. *Research Policy* 49(5), 103957. doi: 10.1016/j.respol.2020.103957.

Chan, J. Y., Chan, K. C., Tong, J. Y., & Zhang, F. F. (2016). Using Google Scholar citations to rank accounting programs: A global perspective. *Review of Quantitative Finance and Accounting*, 47(1), 29–55. doi: 10.1007/s11156-014-0493-x.

Currie, R. R., & G. S. Pandher (2011). Finance journal rankings and tiers: An active scholar assessment methodology. *Journal of Banking & Finance* 35(1), 7–20. doi: 10.1016/j.jbankfin.2010.07.034.

Currie, R. R., & G. S. Pandher (2020). Finance journal rankings: Active scholar assessment revisited. *Journal of Banking & Finance*, 111, 105717. doi: 10.1016/j.jbankfin.2019.105717.

Dai, R., Donohue, L., Drechsler, Q., & Jiang, W. (2021). Dissemination, publication, and impact of finance research: When novelty meets conventionality. *SSRN Working Paper.* doi: 10.2139/ssrn.3803654.

Dominitz, J., & F Manski, C. (2017). More data or better data? A statistical decision problem. *The Review of Economic Studies*, 84(4), 1583–1605. doi: 10.1093/restud/rdx005.

Ederington, L. H. (1979). Aspects of the production of significant financial research. *Journal of Finance*, 34(3), 777–786. doi: 10.2307/2327443.

Hamermesh, D. S. (2013). Six decades of top economics publishing: Who and how? *Journal of Economic Literature*, 51(1), 162–172. doi: 10.1257/jel.51.1.162.

Harvey, C. R., & Hirshleifer, D. A. (2020). Up or out: Resetting norms for peer reviewed publishing in the social sciences. *SSRN Working Paper*. doi: 10.2139/ssrn.3744513.

Heck, J. L., Cooley, P. L., & Hubbard, C. M. (1986). Contributing authors and institutions to the Journal of Finance: 1946–1985. *Journal of Finance*, 41(5), 1129–1140. doi: 10.1111/j.1540-6261.1986.tb02535.x.

Holden, C. W. (2017). Do acceptance and publication times differ across finance journals? *Review of Corporate Finance Studies* 6(1), 102–126. doi: 10.1093/rcfs/cfx011.

Karolyi, A. G. (2011). The ultimate irrelevance proposition in finance? *Financial Review*, 46, 485–512. doi: 10.1111/j.1540-6288.2011.00309.x.

Karolyi, A. G. (2016). Home bias, an academic puzzle. *Review of Finance*, 20(6), 2049–2078. doi: 10.1093/rof/rfw007.

Kerl, A., Miersch, E., & Walter, A. (2018). Evaluation of academic finance conferences. *Journal of Banking & Finance*, 89, 26–38. doi: 10.1016/j.jbankfin.2018.01.014.

Kim, E. H., Morse, A., & Zingales, L. (2009). Are elite universities losing their competitive edge? *Journal of Financial Economics*, 93(3), 353–381. doi: 10.1016/j.jfineco.2008.09.007.

Klemkosky, R. C., & Tuttle, D. L. (1977). The institutional source and concentration of financial research. *Journal of Finance*, 32(3), 901–907. doi: 10.2307/2326321.

Ljungqvist, A., Malloy, C., & Marston, F. (2009). Rewriting history. *Journal of Finance*, 64(4), 1935–1960. doi: 10.1111/j.1540-6261.2009.01484.x.

Mayordomo, S., Peña, J. I., & Schwartz, E. S. (2014). Are all credit default swap databases equal? *European Financial Management*, 20(4), 677–713. doi: 10.1111/j.1468-036X.2013.12023.x.

Netter, J. M., Poulsen, A. B., & Kieser, W. P. (2018). What does it take? Comparison of research standards for promotion in finance. *Journal of Corporate Finance*, 49, 379–387. doi: 10.1016/j.jcorpfin.2018.01.001.

Reinartz, S. J., & Urban, D. (2017). Finance conference quality and publication success: A conference ranking. *Journal of Empirical Finance*, 42, 155–174. doi: 10.1016/j.jempfin.2017.03.001.

Schwert, G. W. (2021). The remarkable growth in financial economics, 1974–2020. *Journal of Financial Economics*, 140, 1008–1046. doi: 10.1016/j.jfineco.2021.03.010.

Smith, S. D. (2004). Is an article in a top journal a top article? *Financial Management* 33(4), 133–149.

Trauner, M. (2017). Buying the haystack: New roles for academic business libraries. *The Academic Business Librarianship Review*, 2(2), 1–4. doi: 10.3998/ticker.16481003.0002.201.

Table 1: Overview of commercial databases and providers

This table lists common databases used in financial research. Panel A covers the most common databases, their providers, and their data coverage. Note that it is difficult to identify which papers rely on WRDS access. Yet, from our experience, almost everybody who is using CRSP access it via WRDS. Panel B lists more specialized databases, their specific use, and their coverage. We use the following mnemonics: US for United States, GL for Global, MF for mutual funds.

Panel A: Most common data providers

		Market data			Accounting Data		
Database	**Provider**	**Stocks**	**Bonds**	**Funds**	**Listed**	**Private**	**Quarterly**
CRSP	CRSP (via WRDS)	US		USMF			
Compustat/Capital IQ	Standard & Poor's	GL	GL		GL		US
Bloomberg	Bloomberg	GL	GL				
Factset	FactSet	GL	GL	GL	GL		
Datastream/Worldscope	Refinitiv	GL	GL		GL		US

Panel B: Further databases for financial research

Data	**Provider**	**Accessible via WRDS**	**Content**	**Coverage**
Osiris	BvD	extra	Accounting	GL
Amadeus	BvD	extra	Accounting	EU
Orbis	BvD	extra	Accounting	GL
BankScope	BvD	extra	Bank	GL
MergentFISD	Mergent	extra	Bonds	GL
Markit		extra	CDS	GL
SDC	Refinitiv	extra	Deals	GL
Zephyr	BvD		Deals	GL
Morningstar Data	Morningstar		Hedge Funds	GL
Sustainalytics	Morningstar	extra	ESG	GL
KLD Stats	MSCI	extra	ESG	GL
Execucomp	Standard & Poor's	extra	Executives	S&P 1500
BoardEX	Euromoney	extra	Executives	GL
Ken French		free	Factors	US/GL
EDGAR	SEC		Filings	US
HFR		extra	Hedge Funds	GL
Lipper Tass	Refinitiv	extra	Hedge Funds	GL
Eurekahedge		extra	Hedge Funds	GL
DealScan	Refinitiv	extra	Loans	GL
IFS	IMF		Macro data	GL
Global Finance Data			Macro data	GL
Yahoo! Finance			Market	GL
Factiva	Dow Jones		news engine	GL
Lexis-Nexis	RELX		news engine	GL
Option Metrics		extra	Options	US/GL
CRSP Mutual Funds	CRSP	extra	Ownership	US
CDA/Spectrum	Refinitiv		Ownership	US
Thomson Financial Insider	Refinitiv	extra	Ownership	US
I/B/E/S	Refinitiv	extra	Stock analysts	GL
CSMAR		extra	Stocks/Accounting	CN
TRACE	FINRA	free	Bonds	US
TAQ	NYSE	extra	Trades/Quotes	NYSE
Nastraq	Nasdaq		Trades/Quotes	Nasdaq
ISSM		extra	Trades/Quotes	US
ISS/RiskMetrics			Votes/Proposals	Russel 3000
WDI			Macro data	GL

Table 2: Empirical focus of journals and database identification

This table reports summary statistics about our identification of financial databases for each of the 16 finance journals considered. *Empirical* refers to the percentage of papers we judge to be empirical following our methodology described in Section 3. *Database* indicates the percentage of papers where we identify at least one database. We report this number separately for all papers published in a journal and for papers that we classified as empirical. *Avg. # of databases* (*authors*) indicates the average number of databases (authors) and standard deviation of how many different databases were identified, conditional on observing at least one data source from our list of 87 databases.

	Journal information			Database information			
Journal name	**n**	**Avg. # authors**	**% empirical**	**% only emp.**	**% all**	**Avg. #**	**Std. Dev.**
European Financial Management	461	2.32	90.89	69.69	65.94	2	1.21
Finance & Stochastic	467	2.05	33.83	12.66	10.49	1.1	0.31
Financial Management	373	2.38	97.59	92.58	91.15	3.09	1.68
Journal of Financial and Quantitative Analysis	728	2.34	91.62	75.71	70.88	3.1	1.79
Journal of Banking & Finance	2,742	2.34	88.04	71.42	66.16	2.15	1.36
Journal of Corporate Finance	894	2.37	93.06	87.5	82.44	3.10	1.69
Journal of Empirical Finance	518	2.31	96.91	71.71	71.62	2.02	1.41
Journal of Finance	1,292	2.28	84.67	83.46	74.46	2.78	1.68
Journal of Financial Economics	1,617	2.35	90.11	89.22	82.93	3.21	1.86
Journal of Financial Intermediation	346	2.12	67.34	66.95	50.00	2.45	1.58
Journal of Financial Markets	368	2.22	86.96	75.63	67.12	2.43	1.45
Journal of International Money and Finance	1,324	2.13	86.77	72.39	65.76	1.68	0.88
Journal of Money, Credit and Banking	772	2.06	83.42	35.56	35.62	1.63	1.01
Mathematical Finance	449	2.06	37.42	12.5	10.02	1.09	0.29
Review of Financial Studies	1,251	2.29	78.42	83.18	70.5	3.04	1.98
Review of Finance	485	2.26	75.26	76.99	63.3	2.51	1.63
Total	14,087	2.27	83.53	74.41	65.55	2.56	1.67

Table 3: Ranking most common databases in financial research

This table lists the most common databases, ranked by the frequency with which they have been used in finance publications between 2000 and 2016. Columns 2 and 3 report the percentage of all/empirical papers that have been mentioning each respective database. Further columns contrast the same percentages from four sample splits against each other: author affiliation with top 10 business school vs others, top 5 finance journals vs others, award-winning research vs other, and international vs US data. * indicates that the database is available without subscriptions or fees.

#	Database	% all	% empirical	Top schools	Other schools	Top 5 journals	Other journals	Award winner	Other	International data	US data
1	CRSP	29.14	34.19	46.49	32.57	51.45	23.26	53.95	34.06	12.39	48.07
2	COMPUSTAT	23.37	27.5	36.26	26.35	40.89	19.02	63.16	27.27	10.84	38.11
3	Datastream	13.49	15.65	10.23	16.36	12.34	17.75	11.84	15.67	26.94	8.46
4	SDC	9.36	11.09	12.35	10.93	16.04	7.96	17.11	11.05	5.46	14.67
5	Bloomberg	8.53	9.45	7.89	9.66	10.28	8.93	10.53	9.44	11.12	8.39
6	IBES	6.24	7.44	8.26	7.33	11.28	5	13.16	7.4	3.72	9.81
7	IFS	5.52	6.22	3.87	6.53	2.21	8.76	3.95	6.24	14.07	1.22
8	Execucomp	4.14	4.85	5.48	4.77	7.43	3.22	7.89	4.83	0.68	7.51
9	ISS	4.22	4.63	5.04	4.58	6	3.76	2.63	4.64	3.1	5.61
10	TAQ	3.89	4.62	3.8	4.73	6.4	3.5	5.26	4.62	1.44	6.65
11	EDGAR	3.76	4.38	5.12	4.28	6.38	3.11	14.47	4.31	1.33	6.31
12	CDASpectrum	3.56	4.22	6.43	3.93	7.43	2.19	9.21	4.19	1.2	6.15
13	WDI	3.63	4.22	3.36	4.33	2.74	5.15	2.63	4.23	9.51	0.85
14	KenFrench*	3.46	4.08	8.33	3.52	7.65	1.82	7.89	4.05	2.34	5.19
15	Worldscope	3.35	3.98	3.73	4.01	5	3.33	6.58	3.96	8.19	1.29
16	Factiva	3.02	3.6	4.17	3.53	5.32	2.51	10.53	3.56	2.36	4.39
17	TRACE	3.21	3.46	3.51	3.45	2.63	3.99	2.63	3.46	3.87	3.2
18	Morningstar	2.5	2.91	4.61	2.68	4.62	1.82	3.95	2.9	1.6	3.74
19	CBOE Volatility Index*	2.48	2.85	3.07	2.82	3.4	2.5	5.26	2.83	2.49	3.07
20	Bureau of Labor Statistics	2.49	2.76	3.22	2.7	3.75	2.14	6.58	2.74	1.55	3.53
21	BankScope	2.43	2.75	0.66	3.03	1.1	3.8	0	2.77	5.83	0.79
22	CensusBureau	2.29	2.6	3.73	2.45	4.16	1.61	6.58	2.57	1.25	3.46
23	WRDS	1.99	2.35	3.58	2.19	3.66	1.53	5.26	2.34	0.79	3.35
24	Dealscan	1.9	2.23	3.73	2.03	3.92	1.15	1.32	2.23	0.79	3.14
25	Sage	1.89	2.05	2.19	2.03	2.06	2.04	5.26	2.03	2.56	1.72
26	CRSPMutualFund	1.59	1.9	3.73	1.66	3.53	0.87	2.63	1.9	0.5	2.8
27	LexisNexis	1.35	1.57	1.68	1.56	1.99	1.31	7.89	1.53	0.9	2
28	CompactDisclosure	1.11	1.33	1.32	1.33	1.86	0.99	3.95	1.31	0.39	1.92
29	Markit	1.02	1.11	1.24	1.1	1.56	0.83	2.63	1.1	0.94	1.22
30	HFR	0.95	1.1	1.97	0.98	1.91	0.58	0	1.1	0.31	1.6

Table 4: Use of databases by topics

This table displays the popularity of common databases depending on which sub-discipline in finance it belongs to. For each topic, we report the percentage of empirical papers making use of a database. Our approach to identify a paper's topic is based on 20 clusters formed by latent Dirichlet allocation (LDA), which we further collapse into 14 financial topics as described in detail in Section 3. We make sure that the table features all 87 databases from our list that make it into the top five of a topic.

Database	Micro-structure	Corporate Finance	Portfolio Mgmt	Option Pricing	Asset Pricing	Financial Econ.	International Finance	Governance	Macro	IPO/ M&A	Banking	Theory	Fixed Income	Equities
CRSP	36.26	49.19	49.59	17.42	45.45	26.68	8.72	35.6	11.8	65.33	9.49	6.73	28.53	78.3
COMPUSTAT	13.85	70.62	12.97	4.55	25.45	14.73	4.36	41.98	5.9	51.19	10.31	3.8	25.94	51.46
Datastream	12.26	11.72	11.67	14.02	33.64	16.47	45.41	11.11	16.77	18.4	15.19	2.34	22.19	19.95
Bloomberg	16.07	5.01	8.43	19.7	15.45	5.97	14.91	4.99	10.87	9.89	10.4	3.65	36.31	7.47
SDC	4.97	19.28	3.57	1.14	0.91	3.54	3.21	13.97	2.38	67.46	3.71	1.46	7.49	6.65
ISS	1.48	6.01	4.05	2.65	2.27	4.93	2.52	13.64	1.45	5.01	1.9	1.46	3.75	1.98
TRACE	5.5	2.85	1.78	1.89	1.36	4.41	3.9	3.19	3.21	2	3.89	1.75	11.82	1.75
IBES	10.04	19.43	2.76	0.76	0.45	3.07	3.44	5.26	0.83	15.89	1.18	0.44	3.17	23.22
Ken French	2.22	3.62	10.53	0.76	10.91	3.71	0.46	1.66	1.35	3.88	0.45	0.73	1.73	19.84
CBOE Volatility Index	2.54	0.93	1.46	20.83	9.55	3.65	4.59	0.13	2.07	0.75	2.08	1.02	7.78	5.37
TAQ	34.14	2.31	1.62	1.89	3.64	2.09	1.61	0.73	0.1	3.13	0.45	0.88	2.02	8.17
IFS	0.42	1.93	0.32	0	8.18	6.96	35.32	2.99	28.36	0.88	6.06	0.88	1.44	0.58
Morningstar	1.06	0.85	35.66	0	4.09	1.51	1.61	0.86	0.41	1.13	0.54	0.58	0.86	2.33
WDI	0.63	3.78	1.3	0	2.27	2.49	17.2	2.33	20.08	1.13	5.88	0.29	0.29	0.47
Execucomp	0.32	8.71	1.3	1.89	0.45	2.49	0	22.02	0.21	5.63	0.54	0.88	0.86	0.58
Markit	1.69	0.69	0.65	1.89	1.36	0.29	0.92	0.33	1.04	0.5	1.54	0.15	12.68	0.47
BankScope	0	0.77	1.62	0	0.45	0.87	0.92	1.26	4.35	0.88	19.26	0	0.29	0.23
CRSP Mutual Fund	0.42	0.46	27.23	0	1.82	0.75	0.69	0.27	0	0.63	0.09	0	0.58	1.63
Total	946	1297	617	264	220	1724	436	1503	966	799	1106	684	347	857

Table 5: Identifying clusters of databases

This table reports the correlation coefficients indicating the overlap of two databases in the same paper. The reported numbers have to be interpreted in a conditional sense, i.e., what is the probability to observe the database reported in a column if we observe the database reported in each row, which is why the reported matrix is not symmetric. To conserve space, we restrict the table to the 20 most common databases in our sample.

NO	Database	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	D13	D14	D15	D16	D17	D18	D19	D20
D1	CRSP	1	0.61	0.1	0.03	0.23	0.08	0.17	0.01	0.11	0.08	0.1	0.09	0.11	0.01	0.1	0.02	0.07	0.06	0.04	0.03
D2	COMPUSTAT	0.76	1	0.08	0.03	0.27	0.07	0.19	0.01	0.16	0.1	0.06	0.11	0.12	0.02	0.08	0.04	0.08	0.02	0.02	0.04
D3	Datastream	0.22	0.13	1	0.05	0.1	0.16	0.07	0.09	0.01	0.05	0.03	0.03	0.02	0.06	0.04	0.15	0.05	0.03	0.05	0.01
D4	TRACE	0.32	0.25	0.23	1	0.11	0.19	0.06	0.05	0.02	0.04	0.06	0.03	0.03	0.02	0.04	0.04	0.04	0.02	0.02	0.02
D5	SDC	0.72	0.68	0.15	0.03	1	0.1	0.19	0.02	0.12	0.09	0.05	0.16	0.1	0.03	0.05	0.08	0.15	0.02	0.01	0.02
D6	Bloomberg	0.29	0.2	0.27	0.07	0.11	1	0.07	0.04	0.02	0.06	0.06	0.06	0.03	0.04	0.04	0.06	0.06	0.04	0.08	0.03
D7	IBES	0.78	0.72	0.15	0.03	0.28	0.09	1	0.01	0.1	0.09	0.14	0.09	0.22	0.03	0.09	0.08	0.11	0.02	0.03	0.03
D8	IFS	0.05	0.06	0.23	0.03	0.03	0.07	0.01	1	0	0.02	0	0.01	0	0.23	0.01	0.05	0.01	0	0.02	0.02
D9	Execucomp	0.78	0.92	0.02	0.01	0.28	0.04	0.15	0.01	1	0.25	0.01	0.15	0.15	0	0.03	0.01	0.12	0.02	0	0.04
D10	ISS	0.56	0.59	0.17	0.03	0.21	0.12	0.15	0.03	0.26	1	0.03	0.12	0.14	0.03	0.04	0.08	0.1	0.03	0.02	0.02
D11	TAQ	0.74	0.35	0.09	0.04	0.11	0.13	0.23	0.01	0.01	0.03	1	0.05	0.14	0.01	0.06	0.01	0.05	0.02	0.05	0.01
D12	EDGAR	0.74	0.66	0.11	0.02	0.41	0.13	0.16	0.01	0.17	0.12	0.05	1	0.15	0.01	0.08	0.04	0.16	0.08	0.01	0.03
D13	CDASpectrum	0.92	0.76	0.08	0.03	0.28	0.08	0.39	0	0.17	0.15	0.15	0.15	1	0.02	0.1	0.03	0.11	0.09	0.01	0.02
D14	WDI	0.07	0.15	0.23	0.01	0.08	0.09	0.05	0.34	0	0.03	0.01	0.01	0.02	1	0.01	0.19	0.03	0.01	0.03	0.01
D15	KenFrench	0.85	0.51	0.14	0.03	0.14	0.08	0.16	0.02	0.04	0.05	0.07	0.09	0.1	0.01	1	0.03	0.05	0.09	0.03	0.02
D16	Worldscope	0.19	0.25	0.58	0.04	0.21	0.13	0.15	0.08	0.01	0.09	0.01	0.05	0.03	0.2	0.03	1	0.08	0.03	0.01	0.01
D17	Factiva	0.66	0.61	0.22	0.04	0.47	0.16	0.23	0.02	0.16	0.13	0.06	0.19	0.13	0.03	0.06	0.09	1	0.04	0.01	0.02
D18	Morningstar	0.71	0.22	0.15	0.02	0.08	0.12	0.05	0.01	0.03	0.05	0.03	0.12	0.13	0.02	0.13	0.04	0.05	1	0.03	0.01
D19	CBOE Volatility Index	0.44	0.21	0.29	0.03	0.06	0.27	0.08	0.04	0.01	0.03	0.08	0.02	0.02	0.04	0.05	0.02	0.01	0.03	1	0.03
D20	Bureau of Labor Statistics	0.4	0.36	0.07	0.03	0.1	0.1	0.07	0.05	0.07	0.04	0.02	0.04	0.03	0.01	0.03	0.01	0.03	0.01	0.03	1

Table 6: Market share of financial data providers

This table restates Table 5 by aggregating all distinct data products offered by the same data provider. Using the most common data vendors considered for financial core products, we report the percentage of empirical papers that employ one of their products. We report this metric separately for the 14 financial topics introduced earlier and presented in Section 3.

Provider	Total	Corp. Finance	Portfo-lio Mgmt	Option Pricing	Asset Pricing	Financial Econ.	Int. Finance	Gov.	Macro	IPO/ M&A	Banking	Theory	Fixed Income	Equi-ties	Micro-structure
Refinitiv	36.04	28.86	52.51	36.63	17.05	37.73	24.88	50.69	35.6	21.95	80.73	26.4	5.56	34.29	51.46
CRSP	34.19	36.26	49.19	49.59	17.42	45.45	26.68	8.72	35.6	11.8	65.33	9.49	6.73	28.53	78.3
Standard & Poor's	28.85	14.8	71.16	15.07	7.58	27.73	15.95	4.82	44.91	7.04	52.82	11.21	4.09	28.53	52.04
Bloomberg	9.45	16.07	5.01	8.43	19.7	15.45	5.97	14.91	4.99	10.87	9.89	10.4	3.65	36.31	7.47
Morningstar	2.93	1.16	0.85	35.66	0	4.09	1.57	1.61	0.86	0.41	1.13	0.63	0.58	0.86	2.33
Bureau van Dijk	2.34	0.11	3.78	0.81	0	0	0.93	1.38	5.12	2.28	2	7.23	0.29	0	0.12
Mergent	1.92	1.69	4.63	0.97	0.76	0.91	0.46	0.92	2.79	0.21	1.75	1.18	0.29	13.26	1.05
Factset	0.65	0.63	1.08	1.13	0	1.36	0.29	1.15	0.86	0.31	0.88	0.45	0	0	0.93

Table 7: Explaining the choice of core data packages

Models 1-5 in this table report the results for a jointly estimated multinomial logit model. The dependent variable is a categorical variable for the five core data packages of Table 1. *Top business school* is defined as 1, if at least one of the authors was affiliated to a top ten business school according to the Financial Times' Global MBA Ranking in the year prior to the publication, and 0 otherwise. *US affiliation* is defined as 1, if at least one the authors is affiliated to a U.S. university or institution, and 0 otherwise. *# authors* is the number of authors of the paper. *International sample* is a binary variable whether the paper works with international data or data outside the US. Model 6 reports a separate estimation of an ordered logit model. In this case the dependent variable is an ordinal variable that ranks the five core data products according to its estimated fixed cost (FactSet=1, Refinitv=2, Bloomberg=3, Compustat/Capital IQ=4, CRSP=5).

	(1) FactSet	(2) Refinitiv	(3) Bloomberg	(4) Compustat	(5) CRSP	(6) Core
Top business school	-0.164	0.059	(base)	0.420*	0.506**	0.278***
	(-0.182)	(0.386)		(1.843)	(2.528)	(3.019)
US affiliation	0.838	-0.423**		1.085***	1.243***	1.098***
	(1.530)	(-2.304)		(7.343)	(9.500)	(7.854)
# authors	0.307	-0.006		-0.125	-0.069	-0.019
	(0.971)	(-0.122)		(-1.637)	(-1.394)	(-0.530)
International sample	0.428	0.644**		-1.133***	-2.007***	-1.950***
	(0.512)	(2.510)		(-5.172)	(-10.351)	(-12.349)
Constant	-5.619***	0.324		0.178	1.687***	
	(-5.030)	(1.252)		(0.647)	(8.410)	
Observations	6,864	6,864	6,864	6,864	6,864	7,096
Cluster	Journal	Journal	Journal	Journal	Journal	Journal

Table 8: Determinants of the number of databases

This table reports results for regression models using the number of databases in a paper as dependent variable. Independent variables are the journal's impact factor varying by year, the log of the number of citations for a given paper, a dummy whether a paper includes robustness tests/sections, a binary variable whether the paper works with international data or data outside the US, as well as topic and year fixed effects. The set of additional control variables closely follows Berninger et al. (2021), including title length, # authors, # references, # tables, # figures, lead articles, article order, special issues, award winners, top business schools, US affiliation, authors' highest number of top 3 publications, top 3 journal editors, outstanding scholars, and the LDA fit with created clusters. We cluster standard errors at the journal level.

	(1)	(2)	(3)	(4)	(5)
Impact factor	0.44***	0.36***	0.20**	0.14**	0.13**
	(0.108)	(0.093)	(0.070)	(0.059)	(0.059)
ln(# cite)			0.20***	0.13***	0.14***
			(0.050)	(0.038)	(0.039)
Robustness tests			0.32***	0.21***	0.20***
			(0.045)	(0.036)	(0.036)
International sample			-0.43***	-0.39***	-0.41***
			(0.103)	(0.088)	(0.092)
Observations	11,215	11,215	11,056	11,056	10,774
R-squared	0.082	0.252	0.286	0.341	0.327
Year FE	No	Yes	Yes	Yes	Yes
Topic FE	No	Yes	Yes	Yes	Yes
Controls	No	No	No	Yes	Yes

Table 9: Databases and its impact on citations

This table reports regression results of publications' average number of citations per year since publication. The dependent variable in Columns 1-4 are based on Crossref citations, and in Columns 5-8 on Google citations, respectively. The two variables of interest are *top databases*, which is a dummy subsuming the most common databases listed in Table 3, and *other databases*, which refers to all other databases identified from our list of 87 databases. The set of additional control variables closely follows Berninger et al. (2021), including title length, # authors, # references, # tables, # figures, lead articles, article order, special issues, award winners, top business schools, US affiliation, authors' highest number of top three publications, top three journal editors, outstanding scholars, and the LDA fit with created clusters. We use robust standard errors.

	Crossref				**Google Scholar**			
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**
Top databases	3.56**	3.44***	1.85***	1.20***	7.04**	6.77***	3.49***	2.18***
	(1.241)	(1.024)	(0.532)	(0.352)	(2.455)	(1.877)	(0.924)	(0.674)
Other databases	1.69***	2.48***	0.99***	0.11	2.79***	4.51***	1.37*	-0.27
	(0.394)	(0.564)	(0.305)	(0.249)	(0.879)	(1.220)	(0.691)	(0.640)
Robustness tests			1.45***	1.01***			2.62***	1.72***
			(0.330)	(0.270)			(0.576)	(0.468)
International sample			0.08	1.08**			0.64	2.50**
			(0.529)	(0.460)			(1.205)	(1.163)
Observations	11,766	11,766	11,766	11,766	11,766	11,766	11,766	11,766
R-squared	0.021	0.063	0.191	0.270	0.015	0.052	0.159	0.215
Controls	No	No	Yes	Yes	No	No	Yes	Yes
Year FE	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Topic FE	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Journal FE	No	No	No	Yes	No	No	No	Yes

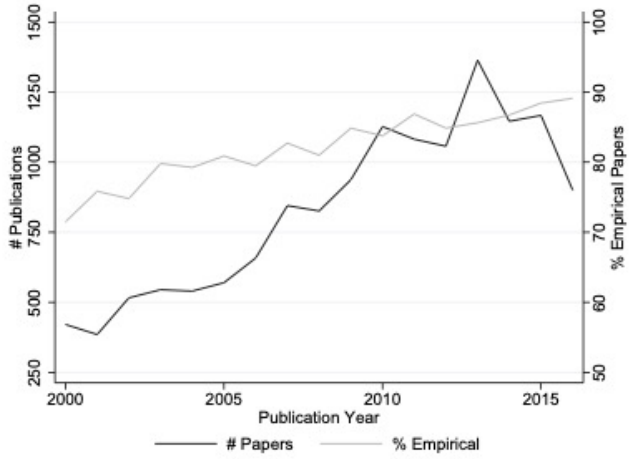
Figure 1: Top 10 keywords and subjects of the LDA clusters

This figure shows the ten keywords with the highest probability for each category and the name of the category using a latent Dirichlet allocation (LDA). The clusters are obtained from Berninger et al. (2021) and extended by labeling 14 different research subjects.

Figure 2: Usage of databases across years

The time-series in Panel A reports the number of papers published per year and the percentage of empirical papers our sample from 2000 to 2016. Using the subsample of empirical papers, Panel B shows the percentage of papers in which we identify at least one database. We also plot the average number of databases identified per year for cases where we have at least one database. In addition, the graph reproduces the same statistic for publications in top five journals and other journals.

Panel A: Ratio of empirical papers per year



Panel B: Average databases per year

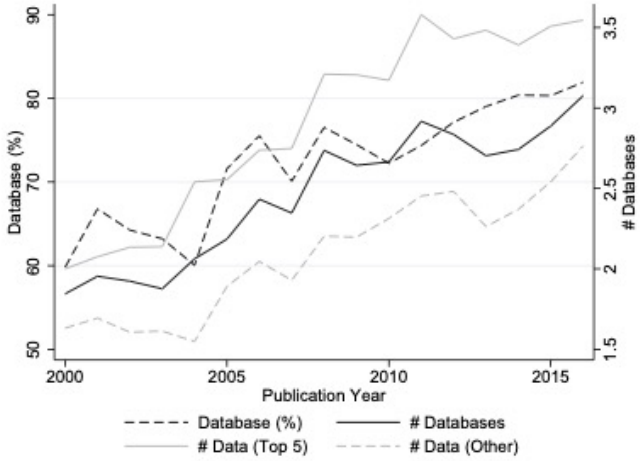
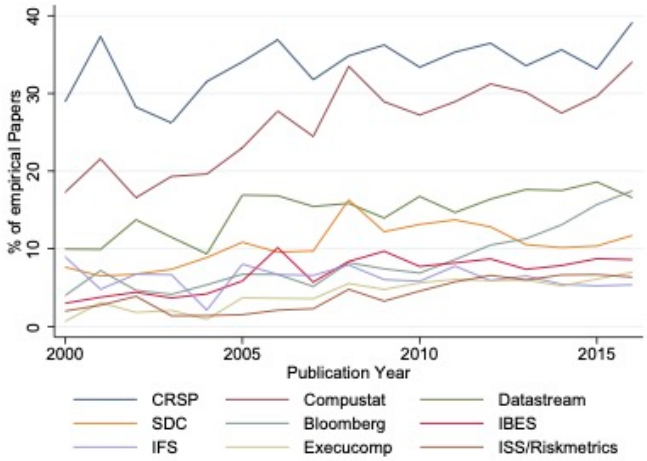


Figure 3: Distribution of database per year

This figure shows the ratios for the 10 most common financial databases between 2000 and 2016.



Appendix A.1: All potential databases

This table extends the list provided in Table 3. The table shows the next common databases, ranked by the frequency with which they have been used in publications between 2000 and 2016. Columns 2 and 3 report the percentage of all/empirical papers that have been mentioning each respective database. Further columns contrast the same percentages from three sample splits against each other: author affiliation with top 10 business school vs others, top 5 finance journals vs others, award-winning research vs other, and international vs US data. * indicates that the databases is available without subscriptions or fees.

#	Database	% all	% empirical	Top schools	Other schools	Top 5 journals	Other journals	Award winner	Other	International data	US data
31	GlobalFinancialData	0.81	0.94	1.68	0.85	1.05	0.87	3.95	0.92	1.51	0.58
32	MergentFISD	0.79	0.93	1.54	0.86	1.67	0.47	2.63	0.92	0.09	1.47
33	CSMAR	0.74	0.88	0.51	0.93	0.46	1.15	0	0.89	2.08	0.13
34	CapitalIQ	0.75	0.87	1.75	0.75	1.53	0.44	1.32	0.86	0.5	1.1
35	Amadeus	0.64	0.76	0.8	0.76	0.59	0.87	1.32	0.76	1.51	0.29
36	ISSM	0.58	0.69	0.51	0.71	1.25	0.33	0	0.69	0.22	0.99
37	BoardEX	0.51	0.61	0.51	0.63	1.07	0.32	1.32	0.61	0.37	0.76
38	FactSet	0.52	0.61	1.24	0.53	0.88	0.44	1.32	0.61	0.92	0.42
39	YahooFinance*	0.52	0.6	0.73	0.59	0.59	0.61	1.32	0.6	0.5	0.67
40	CEX	0.5	0.51	1.02	0.44	0.7	0.39	1.32	0.5	0.26	0.67
41	EIUProducts	0.45	0.5	0.73	0.47	0.46	0.53	0	0.5	1.14	0.1
42	LipperTASS	0.39	0.47	0.66	0.44	1.05	0.1	1.32	0.46	0.13	0.68
43	ThomsonFinancialInsider	0.38	0.46	0.37	0.47	0.74	0.28	0	0.46	0.04	0.72
44	EmergingMarketsDatabase	0.37	0.43	0.73	0.39	0.68	0.28	0	0.44	0.87	0.15
45	NasdaqOMX	0.36	0.41	0.07	0.45	0.46	0.37	2.63	0.39	0.33	0.46
46	Orbis	0.27	0.32	0.15	0.35	0.26	0.36	0	0.33	0.57	0.17
47	Osiris	0.27	0.32	0.37	0.32	0.37	0.29	2.63	0.31	0.66	0.11
48	Preqin	0.23	0.27	0.37	0.26	0.5	0.12	0	0.27	0.11	0.38
49	iPOLL	0.26	0.25	0.07	0.28	0.09	0.36	0	0.26	0.42	0.15
50	LionShares	0.2	0.24	0.51	0.2	0.5	0.07	0	0.24	0.42	0.13
51	Zephyr	0.2	0.24	0.15	0.25	0.26	0.22	0	0.24	0.33	0.18
52	GlobalInsight	0.21	0.23	0.29	0.22	0.26	0.21	2.63	0.21	0.37	0.14
53	Nastraq	0.18	0.22	0.07	0.24	0.26	0.19	0	0.22	0.04	0.33
54	CEIC	0.18	0.21	0.07	0.23	0	0.35	0	0.21	0.52	0.01
55	CCER	0.18	0.2	0.22	0.19	0.09	0.26	0	0.2	0.48	0.01
56	Eventus	0.15	0.18	0.37	0.15	0.2	0.17	0	0.18	0.04	0.26
57	eurofidai	0.16	0.17	0.29	0.15	0.26	0.11	0	0.17	0.15	0.18
58	Euroclear	0.16	0.17	0.07	0.18	0.13	0.19	0	0.17	0.31	0.08
59	MergentOnline	0.13	0.16	0.07	0.17	0.22	0.12	0	0.16	0.22	0.13
60	ICPSR	0.13	0.15	0.37	0.13	0.18	0.14	0	0.15	0.17	0.14
61	BerkeleyOptionsDatabase	0.13	0.15	0.22	0.14	0.28	0.07	0	0.15	0.04	0.22
62	OptionMetrics	0.12	0.14	0.15	0.14	0.24	0.08	0	0.15	0.02	0.22
63	KLDStats	0.1	0.12	0	0.13	0.04	0.17	0	0.12	0	0.19
64	Eurekahedge	0.09	0.11	0.22	0.1	0.2	0.06	0	0.11	0.04	0.15
65	CDP	0.12	0.11	0.22	0.1	0.15	0.08	0	0.11	0.04	0.15

66	RavenPack	0.06	0.07	0.07	0.07	0.11	0.04	0	0.07	0.09	0.06
67	EURIPO	0.06	0.07	0.07	0.07	0.02	0.1	0	0.07	0.13	0.03
68	Mergermarket	0.04	0.05	0	0.06	0.04	0.06	0	0.05	0.02	0.07
69	Sustainalytics	0.03	0.03	0	0.04	0	0.06	0	0.03	0.04	0.03
70	MoodysAnalytics	0.02	0.03	0.07	0.02	0.04	0.01	0	0.03	0.02	0.03
71	IBISWorld	0.01	0.02	0	0.02	0.04	0	0	0.02	0.04	0
72	ThomsonFinancialMA	0.02	0.02	0.07	0.01	0.04	0	0	0.02	0	0.03
73	FAOSTAT	0.01	0.02	0	0.02	0	0.03	0	0.02	0.04	0
74	OneTick	0.01	0.01	0	0.01	0.02	0	0	0.01	0	0.01
75	Eikon	0.01	0.01	0	0.01	0	0.01	0	0.01	0	0.01
76	RepRisk	0	0	0	0	0	0	0	0	0	0
77	Refinitiv	0	0	0	0	0	0	0	0	0	0
78	BankFocus	0	0	0	0	0	0	0	0	0	0
79	QuantQuote	0	0	0	0	0	0	0	0	0	0
80	SocialExplorercom	0	0	0	0	0	0	0	0	0	0
81	BankOne	0.01	0	0	0	0	0	0	0	0	0
82	MergentArchives	0	0	0	0	0	0	0	0	0	0
83	OECDiLibrary	0	0	0	0	0	0	0	0	0	0
84	DataPlanet	0	0	0	0	0	0	0	0	0	0
85	CBInsight	0	0	0	0	0	0	0	0	0	0
86	CEDDS	0	0	0	0	0	0	0	0	0	0
87	DebtWire	0	0	0	0	0	0	0	0	0	0

Appendix A.2: Explaining the choice of research topics

This table shows the results of a multinomial logit model to predict research topics. We use the same independent variables employed in Table 3. *Top business school* is defined as 1, if at least one of the authors was affiliated to a top ten business school according to the Financial Times' Global MBA Ranking in the year prior to the publication, and 0 otherwise. *US affiliation* is defined as 1, if at least one the authors is affiliated to a U.S. university or institution, and 0 otherwise. # *author*s is the number of authors of the paper. *International sample* is a binary variable whether the paper works with international data or data outside the US. Asset pricing is our defined benchmark category, and all coefficients have to be interpreted relative to this benchmark.

	(1) Micro-structure	(2) Corporate Finance	(3) Portfolio Mgmt	(4) Option Pricing	(5) Asset Pricing	(6) Econo-metrics	(7) Interna-tional	(8) Govern-ance	(9) Macro	(10) IPO/ M&A	(11) Banking	(12) Theory	(13) Fixed Income	(14) Equities
Top business school	-0.635***	-0.270	0.003	-0.299		-0.020	-0.087	-0.240	0.152	-0.704***	-0.141	-0.003	-0.045	0.149
	(-3.552)	(-1.319)	(0.022)	(-1.400)		(-0.111)	(-0.296)	(-1.626)	(0.478)	(-3.377)	(-0.594)	(-0.014)	(-0.271)	(0.942)
US affiliation	0.608***	0.758***	0.466**	-0.082		0.027	0.370	0.411*	0.346	0.917***	0.123	-0.331	0.499	0.802***
	(3.142)	(3.264)	(2.469)	(-0.325)		(0.135)	(1.423)	(1.771)	(1.459)	(3.900)	(0.563)	(-1.128)	(1.383)	(3.793)
# authors	0.006	-0.007	-0.000	-0.086		-0.147*	-0.254**	-0.112	-0.271***	0.036	-0.045	-0.362***	-0.055	-0.060
	(0.076)	(-0.065)	(-0.005)	(-0.678)		(-1.700)	(-2.392)	(-1.025)	(-2.597)	(0.361)	(-0.531)	(-4.056)	(-0.446)	(-0.745)
International sample	-0.835**	-0.891***	-1.242***	-1.224***		-0.435	3.184***	-0.645	1.003**	-1.024**	0.178	-1.304***	-0.805**	-0.731**
	(-2.395)	(-2.724)	(-3.910)	(-3.228)		(-1.209)	(3.957)	(-1.594)	(2.171)	(-2.408)	(0.467)	(-4.408)	(-2.272)	(-2.350)
Constant	1.494***	1.718***	1.206***	0.910**		2.589***	-1.430***	2.261***	1.291***	1.097**	1.582***	2.576***	0.634*	1.293***
	(4.454)	(4.527)	(6.550)	(2.490)		(8.930)	(-2.788)	(3.784)	(2.831)	(2.509)	(3.778)	(4.906)	(1.683)	(4.568)
Observations	11,766	11,766	11,766	11,766	11,766	11,766	11,766	11,766	11,766	11,766	11,766	11,766	11,766	11,766
Cluster	Journal	Journal	Journal	Journal	Journal	Journal	Journal	Journal	Journal	Journal	Journal	Journal	Journal	Journal

Appendix A.3: Detailed analysis of databases and citations

This table extends the findings of Table 8 and reports regression results of publications' average number of citations per year since publication. The dependent variable in Columns 1-4 are based on Crossref citations, and in Columns 5-8 on Google citations, respectively. We create a binary variable for each of the 20 most common databases. The set of additional control variables closely follows Berninger et al. (2021), including title length, # authors, # references, # tables, # figures, lead articles, article order, special issues, award winners, top business schools, US affiliation, authors' highest number of top 3 publications, top 3 journal editors, outstanding scholars, and the LDA fit with created clusters. We use robust standard errors.

	(1) All	(2) All	(3) All	(4) Top3	(5) All	(6) Top3
CRSP	0.20	0.29	-0.44	-0.34	-1.90**	-2.96**
	(0.366)	(0.370)	(0.352)	(0.721)	(0.739)	(1.467)
COMPUSTAT	1.70***	1.92***	1.22***	2.37**	2.84***	5.55***
	(0.469)	(0.421)	(0.454)	(0.987)	(0.965)	(2.039)
Datastream	0.33	0.29	0.53**	0.94	1.24*	2.80
	(0.281)	(0.277)	(0.264)	(0.963)	(0.653)	(2.407)
TRACE	0.40	0.33	0.64	2.29	0.48	1.94
	(0.625)	(0.625)	(0.582)	(2.027)	(1.184)	(4.167)
SDC	-0.77*	-1.80***	-1.26***	-1.80**	-2.28**	-2.84
	(0.428)	(0.384)	(0.398)	(0.807)	(0.932)	(1.906)
Bloomberg	0.81**	0.51	0.54*	0.93	0.74	1.26
	(0.340)	(0.344)	(0.324)	(0.911)	(0.751)	(2.029)
IBES	-0.03	-0.09	-0.12	0.47	-0.85	0.39
	(0.465)	(0.468)	(0.429)	(0.883)	(0.923)	(1.921)
IFS	-1.03***	-0.96***	-0.08	3.41	0.85	8.16
	(0.370)	(0.344)	(0.360)	(2.332)	(0.879)	(5.934)
Execucomp	1.46**	1.80***	0.59	0.39	1.14	1.35
	(0.642)	(0.633)	(0.607)	(1.216)	(1.436)	(2.721)
ISS	0.74	0.89	0.80	1.24	1.60	3.83
	(0.575)	(0.578)	(0.543)	(1.264)	(1.270)	(3.031)
TAQ	0.06	-0.96*	0.07	0.53	0.27	0.63
	(0.551)	(0.548)	(0.520)	(1.142)	(1.152)	(2.514)
EDGAR	0.02	-0.24	-0.21	-0.00	-1.25	-0.97
	(0.553)	(0.557)	(0.518)	(1.048)	(1.086)	(2.193)
CDASpectrum	0.95	0.87	0.39	0.62	1.28	2.53
	(0.653)	(0.652)	(0.608)	(1.082)	(1.403)	(2.526)
WDI	1.09**	1.11**	1.03**	1.87	2.58**	4.95
	(0.550)	(0.546)	(0.519)	(2.077)	(1.293)	(5.240)
KenFrench	3.04***	3.47***	1.80**	2.23*	3.48*	4.27
	(0.840)	(0.827)	(0.814)	(1.249)	(1.810)	(2.833)
Worldscope	4.45***	4.55***	3.32***	6.53***	7.07***	12.89***
	(0.816)	(0.835)	(0.763)	(1.932)	(1.747)	(4.601)
Factiva	0.53	0.35	0.19	1.42	-0.04	1.90
	(0.613)	(0.614)	(0.558)	(1.273)	(1.194)	(2.742)
Morningstar	-1.58***	-2.10***	-2.51***	-3.00***	-5.94***	-8.30***
	(0.556)	(0.554)	(0.514)	(1.122)	(0.976)	(2.125)
CBOEVolatilityIndex	2.79***	2.77***	2.32***	4.12**	5.17***	8.41**
	(0.772)	(0.759)	(0.733)	(1.783)	(1.696)	(3.451)
BureauofLaborStatistics	-0.13	0.24	-0.71	-1.11	-2.04	-2.92
	(0.745)	(0.746)	(0.708)	(1.649)	(1.523)	(3.565)
BankScope	4.17***	5.25***	4.09***	5.73	8.75***	11.13
	(0.655)	(0.649)	(0.614)	(3.580)	(1.352)	(7.915)
other	0.92***	0.92***	0.17	0.29	-0.09	0.37
	(0.276)	(0.277)	(0.262)	(0.666)	(0.605)	(1.513)
Observations	11,766	11,766	11,766	3,532	11,766	3,532
R-squared	0.199	0.190	0.276	0.166	0.222	0.158
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Topic FE	No	Yes	Yes	Yes	Yes	Yes
Journal FE	No	No	Yes	Yes	Yes	Yes