Information Intermediaries and the Distorting Effect of Incomplete Data

Sara Easterwood [*]

May 7, 2024

Abstract

Financial data vendors intermediate the flow of information from firms to investors. I study frictions that arise in the context of this intermediation by focusing on one of the most prominent data vendors in the finance industry: Standard & Poor's ('S&P') Compustat database. Compustat provides subscribers with decades of 10-K and 10-Q data; however, it does not cover every public firm in every period. I show that a significant fraction of institutional investors do not invest in firms with missing data – institutional ownership is over 36% below its unconditional mean for firms not covered in Compustat. A policy change instituted at S&P in the early 1990s provides a quasi-natural experiment to confirm a plausibly causal association between Compustat data coverage and institutional investor demand. In a battery of empirical tests, I then show that limited access to financial data is associated with lower informational efficiency of equity prices. This highlights the role that data vendors play in facilitating the flow of information within financial markets.

[*] Easterwood, saraeast@vt.edu. I am grateful to my advisors Bradley Paye (chair), Roger Edelen, Gregory Kadlec, and Jeffry Netter for their invaluable guidance and support. I thank Tyler Beason, Elizabeth Bickmore, Felipe Cabezón, Robert Davidson, Raman Kumar, Andrew MacKinlay, George Morgan, Alex Pecora, Vijay Singal, Wei-ling Song, Yessenia Tellez, Ngoc-Khanh Tran, Jin Xu, Penfei Ye, and seminar participants at Louisiana State University and Virginia Tech for helpful comments. I thank Randall Billingsley, Chris Brightman, and Michael Kender for providing incite into the sources and uses of financial data in the professional asset management industry. I thank all members of S&P Global's staff who provided invaluable information regarding the evolution of the Compustat database over time. All errors are my own.

1 Introduction

A variety of frictions limit the informational efficiency of financial markets. Seminal theoretical work highlights costly information acquisition and processing as among the most salient of such frictions. For example, Grossman and Stiglitz (1980) argue that the higher the cost of obtaining information, the fewer informed investors there will be, and therefore the less informative prices will be. Likewise, Merton (1987) emphasizes that rational investors will only trade certain securities or strategies if the potential gains sufficiently outweigh the information acquisition and related implementation costs.[1] Despite a robust theoretical literature, however, empirically measuring the marginal costs and benefits of becoming informed is challenging. In this paper, I propose a novel empirical setting that tackles this challenge. I use this setting to directly evaluate the influence that information availability can have on investor demand, and how limited access to information affects market efficiency.

The costly nature of information acquisition and processing has given rise to an entire industry of professional data aggregators. These data vendors act as information intermediaries by collecting and aggregating data on clients' behalf. Standard & Poor's ('S&P') Compustat database is one of the oldest and most prominent data vendors in the finance industry, and Compustat has contributed to massive reductions in the collective cost of aggregating and processing public firms' accounting information. However, the database is not and has never been comprehensive. If a significant fraction of market participants rely on Compustat to access firms' accounting information, then they are implicitly relying on the quality and completeness of the data vendor's coverage. I examine how the completeness of Compustat's data coverage affects institutional investor demand, and how limited data coverage can ultimately affect the informational efficiency of equity prices. In doing so, I shed light on the important role that data vendors play in facilitating the flow of information within the economy.

[1] Goldstein and Yang (2017), Blankespoor et al. (2020), and Kothari et al. (2023) provide more complete reviews of the theoretical literature examining information disclosure and acquisition in financial markets.

In research and in practice, it is widely acknowledged that firms' financial statements provide information that is critically important to investors.[2] The importance of this information can arise from many alternative avenues, such as trading strategies that utilize accounting-based firm characteristics or investment mandates that incorporate restrictions based on financial ratios. However, there are substantial costs associated with obtaining and processing financial statement information for a comprehensive selection of firms (Blankespoor et al., 2020; Kothari et al., 2023). Standard & Poor's Compustat database plays a significant role in this regard because it aggregates 10-K and 10-Q filings, and provides clients with a standardized data set consisting of income statement, balance sheet, and statement of cash flow information. Data vendors such as Compustat thus intermediate the flow of financial statement information from firms to (many) investors.[3]

As of 2024, the Compustat database covers tens of thousands of firms over a nearly 75-year historical period. It does not, however, cover every publicly traded firm in all fiscal years or quarters, nor does it provide the same set of financial statement information for all firms at all points in time. Using a comprehensive selection of firm characteristics from the accounting and finance literatures, I first show that missing data (or 'missingness') is a pervasive problem in Compustat that affects firms of all sizes. I then show that firms with no Compustat data coverage have 1.5% lower institutional ownership on average relative to firms that are covered in the database. Given that institutions are the dominate investors in capital markets,[4] and that the average fraction of institutions that own shares of a firm is 4% (standard deviation = 7%), this is an economically significant effect: institutional ownership

[2] For example, the Securities and Exchange Commission states that they require firms to disclose their financial statements so that investors have "the timely, accurate, and complete information they need to make confident and informed decisions about when or where to invest," (https://www.sec.gov/about/what-we-do). Likewise, academic research such as Bushee and Noe (2000), Bushee et al. (2003), and Bird and Karolyi (2016) highlights the importance of financial statement disclosure.

[3] The following quote from a letter written by Pricewaterhouse Coopers and sent to the SEC in 2006 confirms the relevance of data vendors within the finance industry: "Based on our discussions with investors and analysts, we understand that investors acquire the large majority of relevant information, including financial data, from sources not controlled by the reporting entity," and, in fact, "the majority of analytical source material is obtained from data aggregators." (Pricewaterhouse Coopers, LLP, 2006, June 8)

[4] See, e.g., Gompers and Metrick (2001), Badrinath and Wahal (2002), Gutierrez and Kelley (2009), Dasgupta et al. (2011), Edelen et al. (2016), Koijen and Yogo (2019), and many others.

of firms with missing Compustat data is over 36% below the unconditional mean.

I use a quasi-natural experiment to establish plausibly causal evidence of the relation between missing Compustat data and institutional ownership. Prior to 1993, Standard & Poor's collected financial statement data for only a small subset of banks and financial services institutions. In the early 1990s, S&P instituted a policy change which stipulated the following: First, they began collecting financial statement data for all financial services firms from the 1993 fiscal year onward; this data became comprehensively available in the Compustat North America database near the end of 1994. Second, they began back-filling financial statement data for many previously uncovered financial firms, which was released primarily between 1993 and 1994. This policy was triggered by an improvement in the technologies used to collect and process data at Standard & Poor's, and led to a discrete, precipitous reduction in missingness within the Compustat database for financial firms. This policy was unrelated to changes in individual firms' 10-K and 10-Q filings. Using a difference-in-differences analysis, I show that average institutional ownership of treated financial firms is nearly 40% lower than institutional ownership of untreated firms prior to the change in data coverage.[5] This gap in ownership closes following the increase in coverage, and from the late 1990s onward, treated and untreated firms have similar average ownership.

The results associated with the difference-in-differences analysis indicate that a significant fraction of institutions do not invest in firms with missing data in Compustat. This confirms the relevance of Compustat as an information intermediary. In addition, it highlights a surprising quality of the professional asset management industry: even though it is possible to supplement Compustat with self-collected data, the empirical results suggest that many institutions do not do so. I evaluate several potential explanations for why this is the case.

First, institutional investors manage assets on clients' behalf, and there are many agency conflicts that arise because of this principal-agent relation (Lakonishok et al., 1992). In-

[5] To be more specific, average institutional ownership was approximately 3.6-5% for untreated firms in the late 1980s and early 1990s. Average ownership for treated financial firms was approximately 1.5% lower during this time, corresponding to a difference in magnitude of 30–40%.

vestment mandates and regulatory constraints are two ways in which these agency costs are mitigated. Both of these factors plausibly influence institutions' incentive to rely exclusively on Compustat. Specifically, investment mandates broadly specify funds' investment strategies and investable assets. Only those institutions governed by mandates which incorporate or allow for the use of accounting data should be directly affected by Compustat's data coverage. Additionally, as fiduciaries, institutional investors have a legal obligation of due diligence. These due diligence requirements may also impact portfolio managers' incentive to rely on a data vendor. For example, portfolio managers may require data for various financial ratios so that they can clearly communicate to stakeholders why they made certain investment decisions, and why those investments satisfy the relevant prudence requirements. They may be reluctant to self-collect data because this adds the additional burden of defending the integrity and comparability of their self-collected data, relative to data obtained from a well-established data vendor. Empirical analyses support these hypotheses, and suggest that both investment mandates and due diligence constraints influence institutional investors' incentive to self-collect data.

Second, the explicit costs associated with collecting and processing financial statement data may incentivize some portfolio managers to rely exclusively on a data vendor. From an equilibrium perspective, it is only optimal for portfolio managers to self-collect data if the marginal benefits of obtaining the information are greater than the marginal costs associated with collecting that information (Grossman and Stiglitz, 1980). It is possible that, for many institutions, the explicit costs associated with acquiring firms' disclosures and maintaining a database of information for these uncovered firms exceed the benefits they might accrue from obtaining and trading on the additional accounting data. Empirical analyses ultimately provide support for this hypothesis as well. Collectively, this suggests that investment mandates, due diligence constraints, and the explicit costs associated with self-collecting data all influence institutional investors' incentive to rely exclusively on Compustat as an information intermediary.

I conclude this study by evaluating whether limited access to financial statement information affects information assimilation and market efficiency. There are two potential channels through which this may occur: First, financial statement data may inform investors' forecasts of and reactions to news releases, such as earnings announcements. If this is the case, investors' ability to forecast and/or accurately react to information that is released may be limited when they do not have access to prior accounting data. Second, firms not covered in Compustat may face substantially less scrutiny by many market participants who require accounting information (e.g., for due diligence reasons) before investing. If this is the case then, in the spirit of Merton (1987), these uncovered firms' equity prices will be significantly less informationally efficient because of limited investor attention.

In a battery of empirical tests, I find that earnings surprises, post earnings announcement drift, several measures of price delay from Hou and Moskowitz (2005), and several measures of daily return autocorrelations are all significantly larger in magnitude for firms with missing Compustat data. I also find that this effect is mitigated if there are sufficiently many institutions investing in the uncovered firms and/or if there are sufficiently many analysts following the uncovered firms. For example, results indicate that earnings surprises, measured via cumulative abnormal returns around earnings announcements, tend to be 0.3–0.5% larger in magnitude for firms not covered in Compustat. This effect is negated by an (approximately) one standard deviation increase in institutional ownership or analyst coverage. These results are collectively consistent with the notion that limited access to financial statement data reduces the informational efficiency of equity prices via its impact on market participation and investor attention.

Compustat's data coverage has consistently improved over time, and many of the issues that plagued the database in past decades have attenuated. In addition, there are many more alternative sources from which investors can obtain financial statement data post-2010 relative to earlier decades. Firm's electronic filings are now available via the Electronic Data Gathering, Analysis, and Retrieval (EDGAR) and eXtensible Business Reporting Language

(XBRL) databases, and data aggregators such as Bloomberg provide products that directly compete with Compustat. This suggests that frictions related to the intermediation of *financial statement* information have ameliorated over time. However, financial statement data is only one subset of potentially relevant information. The last several decades were accompanied by continuous and exponential improvements in information technologies and data gathering methods. Data vendors such as Glassdoor, the Carbon Disclosure Project, the Privacy Rights Clearinghouse, StockTwits, and other social media platforms now provide information related to employee satisfaction, pollution, cybersecurity risk, retail investor sentiment, and other ESG-related firm characteristics. All of this information is plausibly relevant to a significant fraction of investors. As such, the role that data vendors play as information intermediaries will continue to be relevant in studies of financial markets.

Related Literature

This paper is closely related to an emerging literature which empirically examines the importance of information technologies in capital markets. Gao and Huang (2020) exploit the staggered timing of the implementation of the EDGAR system to show that positive shocks to information dissemination technologies improve the informational efficiency of stock prices. Kim et al. (2023) use the staggered implementation of EDGAR to study the impact that information acquisition costs have on the performance of accounting-based anomalies. They show that average anomaly alphas decline significantly following EDGAR's launch, which suggests that information costs are as relevant as transaction costs and other limits to arbitrage in explaining anomaly returns.

In contrast to these studies, I focus on the frictions that arise in financial markets because of variation in Compustat's data coverage. This remains an important issue because, even after EDGAR's implementation, collecting and aggregating financial statement data is costly enough that many investors continue to turn to professional data aggregators to obtain this information. In addition, as information technologies continue to improve and more potentially relevant data becomes available via data aggregators, the frictions that arise in the context of this information intermediation will continue to be relevant in studies of capital markets.

Several other studies empirically evaluate the role that information technologies play in capital markets. D'Souza et al. (2010) examine the relation between institutional demand and the speed with which accounting information is disseminated via Compustat. Their analyses suggest that institutions prefer richer information environments. Likewise, Da et al. (2011) develop a measure of retail investor attention using Google search activity and Ben-Rephael et al. (2017) develop a measure of institutional investor attention using institutions' news searching activity on Bloomberg terminals. The results in Ben-Rephael et al. (2017) suggest that increased institutional investor attention facilitates information assimilation in equity markets. Akbas et al. (2018) link the delay with which analysts' earnings forecasts are activated in I/B/E/S to measures of investor demand and market efficiency. Schaub (2018) links information dissemination speeds in the First Call database to measures of market efficiency following earnings announcements. Farboodi et al. (2022) develop a measure of the quantity of data investors have about different groups of assets. Bowles et al. (2023) study the timing of anomaly returns around information releases in Compustat. The results in Bowles et al. (2023) are consistent with the notion that delayed information processing by investors at least partially explains anomaly returns.

Recent studies also link investors' information acquisition practices to their performance. Crane et al. (2023) obtain data regarding information requests from EDGAR, and study the information acquisition behavior of hedge fund managers. They show that hedge funds' performance is strongly positively related to the extent of their information acquisition. Likewise, Bowles and Reed (2024) use data regarding information requests from EDGAR to study the information acquisition behavior of mutual fund managers. They show that managers tend to acquire more information about their short positions, and that their short positions tend to generate better returns than their long positions.

This paper also contributes to a growing literature that recognizes the importance of missing and/or incorrect data in prominent economic databases. Chen et al. (2015) use the number of missing variables in Compustat to measure the level of detail in firms' annual 10-K's.[6] Bryzgalova et al. (2023), Chen and McCoy (2023), and Freyberger et al. (2023) each discuss alternative econometric methods which researchers can use to address missing data, and focus primarily on asset pricing applications of their empirical methods. Chychyla and Kogan (2015), Boritz and No (2020), and Du et al. (2023) compare "as-filed" XBRL data to accounting data obtained from Compustat, and emphasize that there are often significant discrepancies between the two. Ljungqvist et al. (2009), Chuk et al. (2013), and Karpoff et al. (2017) each examine the accuracy and completeness of data obtained from I/B/E/S, First Call, and popular financial misconduct databases, respectively.

Finally, this paper is related to an extensive empirical literature that examines frictions in financial markets. Many studies evaluate frictions related to transactions costs (e.g., Hasbrouck, 1991; Lesmond et al., 2004; Frazzini et al., 2014; Novy-Marx and Velikov, 2016; Detzel et al., 2023, etc.) and short-sale constraints (e.g., Shleifer and Vishny, 1997; Jones and Lamont, 2002; Nagel, 2005; Chu et al., 2020, etc.). Relatively few attempt to directly address frictions related to the cost of acquiring and processing information.

2 Financial Statement Information Intermediaries

2.1 Data Vendors

Financial statement information can be important to investors for many reasons. The costly nature of acquiring and processing this information for a comprehensive sample of firms has given rise to a variety of data aggregators who collect and standardize 10-K and 10-Q data.

[6] Notably, Chen et al. (2015) focus on variation in the number of missing items across firms with at least some Compustat data coverage. Their measure of disclosure quality is designed to capture variation in the 'fineness' of firms' disclosures. In contrast, I focus primarily on whether or not a firm is covered in Compustat at all. This is unrelated to the content of firms' SEC filings. Hoitash and Hoitash (2018) propose a refinement of the Chen et al. (2015) measure that is based on the number of accounting items from XBRL filings (equivalently, the number of XBRL tags). Johnston et al. (2024) proposed a further refinement that is based on the number of balance sheet and income statement line items from XBRL filings.

Standard & Poor's is one such data vendor, and Compustat is their database. S&P began developing and selling subscriptions to Compustat in the early 1960s, and Compustat covers balance sheet and income statement information dating back to 1950 for some firms. The fact that Standard & Poor's has continued to offer Compustat subscriptions over the past six decades provides a strong signal regarding the database's success in both industry and academia. Studies in the accounting and information systems literatures cite Compustat as one of the most prevalent financial databases (Chychyla and Kogan, 2015) and in S&P's own words, they are "the global standard in providing critical financial information."[7]

Compustat is not the only accounting statement database available to investors. Other data vendors that provide (or have provided) financial statement information include Compact Disclosure, Dialog, Value Line, and Bloomberg. Many of these alternative databases cover(ed) only a small subset of firms, and coverage criteria is typically based on firm size. As a concrete example, Value Line's financial statement database covers only 1,650 companies (Kim et al., 2023), which corresponds to less than half of all public U.S. firms.

Bloomberg's 'Global Companies Financial Data' database provides 10-K and 10-Q data for public companies in the U.S. and in other countries, and may currently be Compustat's biggest competitor. However, Bloomberg was not founded until the early 1980s, approximately 20 years after S&P began developing Compustat, and anecdotal evidence suggests that it took time for the Bloomberg Terminals to gain popularity. For example, in 1992, James P. Love filed a petition to the SEC outlining why the public should have access to the EDGAR system (Love, 1992, January 14). While data vendors such as S&P Global (formerly The McGraw Hill Company) are explicitly discussed, Bloomberg is not. Likewise, a New York Times article (Barringer and Fabrikant, March 21, 1999) discusses Bloomberg's 'coming-of-age' in the late 1990s, suggesting that, going forward, the company would likely marginalize most other data vendors in the multi-billion dollar financial data industry. In addition, similar to other financial data vendors, Bloomberg's database does not comprehen-

[7] https://www.spglobal.com/en/who-we-are/our-history#fourth

sively cover all public companies. In a private conversation with a member of Bloomberg's Equities Help Desk in November, 2023, they indicated that Bloomberg's coverage was limited to only the largest firms when the database was originally built, and has since expanded over time to cover a broader and more representative sample. They explicitly stated that Bloomberg does *not* currently cover all U.S. firms.

2.2 SEC Resources

Investors can also obtain financial statement information directly from the SEC. There are reference rooms located in Washington D.C., New York, and Chicago, which provide paper copies of firm's financial statements. Studies such as Blankespoor et al. (2020), Gao and Huang (2020), Bowles et al. (2023), Kim et al. (2023), and Kothari et al. (2023) discuss the many issues associated with using these rooms as a source of information. Not only do investors have to be physically present to obtain the information, but there is also evidence suggesting that paper files are routinely lost and/or stolen (Noble, 1982).

In the mid-1990s, the SEC introduced the EDGAR database, where all public corporations are required to electronically file their public disclosures. While several studies have highlighted the role that EDGAR played in massively reducing information acquisition costs (e.g., Gao and Huang, 2020; Kim et al., 2023), others have noted that EDGAR is *not* most investors' primary source of public disclosures (e.g., Pricewaterhouse Coopers, LLP, 2006, June 8; Drake et al., 2015; Blankespoor et al., 2020). In 2009, the SEC began to require all public firms to file their financial statements in eXtensible Business Reporting Language format. The XBRL database was implemented in an effort to facilitate data retrieval and analysis (SEC Release No. 33-9002).

While both XBRL and EDGAR certainly improved investors' ability to *access* information, they have not necessarily improved investors' ability to *process* that information. Firms have considerable within-GAAP discretion over how to structure their financial statements. As such, processing financial statement information for a comprehensive selection of firms requires some form of meaningful standardization. Anecdotal evidence indicates that this is very much a nontrivial task. For example, in a letter to the SEC, Pricewaterhouse Coopers highlights the very high costs that analysts and investors face if they are forced to manually process paper filings, and indicate that this contributes to wide spread reliance on third party intermediaries (Pricewaterhouse Coopers, LLP, 2006, June 8). Likewise, Harris and Morsfiled (2012) point out that, unless both the Financial Accounting Standards Board ('FASB') and the SEC make significant efforts to simplify the underlying taxonomy of financial statements, improving data access and quality via systems such as EDGAR and XBRL is highly unlikely to be sufficient for investors to readily use the data provided.

2.3 Implications

Whether or not a significant fraction of investors have historically relied on Compustat is ultimately an empirical question. Although there are a number of alternative options, none of them is perfectly efficient or cost-less. The remainder of the paper thus focuses on the following: does Compustat data coverage affect investor demand? If so, what are the economic consequences of limited access to financial data for firms not covered in Compustat?

3 Sample Construction

I obtain monthly and daily stock return and price information from CRSP. Data regarding annual and quarterly financial statement information is from the Compustat North America database. The CRSP and Compustat samples cover the period 1962–2022. I match accounting data from the $t - 1$ fiscal year to price information from July in year t through June in year $t + 1$. Additional data regarding the timing of information releases within Compustat North America is from the Compustat Point-In-Time (PIT) database. The PIT data is available for a limited historical period beginning in December 1986.

Institutions' stock holdings data are from the Thomson Reuters 13f database (s34 file).[8]

[8] The Securities and Exchange Commission requires all institutional investors who manage equity investments exceeding $100 million in any of the past four quarters to report their quarterly holdings on form 13F

This data is available at the quarterly frequency beginning in 1980, and contains long-only equity positions for institutional investors with at least \$100 million in total equity under management ('EUM'). I categorize institutional investors using classifications from Brian Bushee's website and from Ralph Koijen's website.[9] The adjusted investor types include: insurance companies; banks; pension funds; mutual fund companies; investment companies/advisors, including hedge funds; and miscellaneous. For robustness, I also obtain data regarding mutual fund holdings from the Thomson Reuters 13f database (s12 file).[10] Finally, I obtain summary data describing analyst coverage and earnings announcements from I/B/E/S, which is available beginning in 1976. The final CRSP/Compustat/13f/IBES merged data set, including the institutional investor type classifications, covers the period January 1980 - December 2021.

I construct two measures of aggregate institutional ownership. The first, denoted $FNIO_{i,q}$, is equal to the number of institutions that hold shares of stock i in quarter q, scaled by the total number of institutions in the 13f data set in quarter q. The second, denoted $FSIO_{i,q}$, is equal to the fraction of stock i's shares outstanding held by institutions in q.[11] Measures of mutual fund ownership, $FNMF$ and $FSMF$, are defined similar to $FNIO$ and $FSIO$, respectively, except using the s12 mutual fund holdings data. Analyst coverage is defined as the total number of analysts covering a firm, and is equal to the number of quarterly earnings forecasts made by unique analysts.

Table 1 reports summary statistics for the institutional holdings data, mutual fund holdings data, and analyst coverage data. The average $FNIO$ has increased over time, from approximately 3.6% in the 1980s to approximately 4.6% in the 2010s. Similarly, the average

within 45 days of the end of the quarter. Holdings of less than 10,000 shares or \$200,000 in market value are exempted.

[9] https://accounting-faculty.wharton.upenn.edu/bushee/ and https://www.koijen.net/index.html, respectively.

[10] It is important to note that a 'Mutual Fund Company' from the s34 file reflects a family of potentially many mutual funds, while a mutual fund from the s12 file reflects a single mutual fund. Koijen and Yogo (2019) categorize institutions from the s34 file as 'Mutual Fund Companies' if 1) their type code is 3, 4, or 5, and 2) their name and assigned number match a record from the s12 file.

[11] Following Lewellen (2011), observations where the fraction of shares outstanding held by institutions exceeds 100% are truncated at 100%. This occurs in approximately 2.5% of cases.

$FSIO$ has increased over time, from approximately 16% in the 1980s to approximately 58% in the 2010s. These summary statistics are consistent with other many studies (e.g., Hong et al., 2000; Gompers and Metrick, 2001; Edelen et al., 2022).

4 Missing Data in Compustat

The Compustat database covers financial statement information for tens of thousands of firms, and is a leading data provider in industry and academia. However, Compustat's data coverage is not comprehensive. Panel A of Figure 1 reports the fraction of firm-month observations with missing characteristic values over time.[12] Results show that many characteristics are missing for over 50% of firms in the 1960s and 1970s. The fraction of firms with missing characteristic data subsequently declines from the mid-1970s through the end of 2020. Characteristics such as asset growth ('a_growth') are available for nearly 100% of firms by the early 2000s. For other characteristics, however, missing data remains a pervasive and non-trivial problem even over the most recent two decades. As a concrete example, the operating profitability characteristic ('op') underlying the Fama and French (2015) profitability factor is missing for over 35% of public firms between 2010 and 2020.[13]

Missing data in Compustat can be traced to two basic sources. First, Compustat does not provide any data coverage for some firm-fiscal-periods. Second, among the firm-fiscal-periods that Compustat covers, the database does not provide the same set of financial statement information for all firms. Accounting variables such as total assets, net income, and other bottom-line items from the balance sheet and income statement are typically only missing if Compustat does not cover a firm in a given period. Other items, such as different types of expenses (SG&A, R&D), can be missing a variety of alternative reasons.

[12] The characteristics, abbreviations, and definitions are reported in Appendix Table 1.

[13] This statistic defers from a similar statistic reported in Bryzgalova et al. (2023), who find that 14% of firms have missing operating profitability in 2020. I believe the difference is related to the treatment of the deferred taxes and investment tax credit (TXDITC) variable. It is common to fill missing values of TXDITC with zeros when constructing book equity and the Fama and French (1992) book-to-market factor (e.g., see Drechsler, 2023). Missing TXDITC does not generally reflect a case where the firm had trivial deferred taxes and investment tax credits and, instead, most often reflects a case where Compustat did not record one or both of these values. TXDITC is missing for around 20% of firms throughout the 2010s.

Whether or not Compustat covers a firm is typically a function of exchange listing, industry membership, and/or time since IPO. The extent to which these factors affect coverage is strongly time varying: in the 1960s, 1970s, and early 1980s, exchange listing was a dominant factor; from the 1960s through the 1990s, industry membership was also an important factor; from the mid-1990s onward, firm age is the dominant explanation. Among individual accounting items, the reasons why data may be missing varying greatly, and can be a function of Compustat's data collection and processing procedures, as well as the underlying structure of individual firms' financial statements. In a companion paper, Easterwood (2024) provides a much more detailed discussion of when and why data is missing in Compustat.

Panel B of Figure 1 reports the fraction of firm-month observations over time with no Compustat coverage. Results show that around 20% (45%) of public firms had no annual (quarterly) data coverage in the late 1970s and 1980s. These uncovered firms are primarily NASDAQ firms and/or firms in the financial services industry. Results also show that Compustat's coverage has improved significantly over time. By the early 2000's, Compustat provides some data for close to 100% of firms. This explains why characteristics such as asset growth and return on assets are non-missing for nearly 100% of firms post-2000: these characteristics can essentially always be constructed if a firm has any data coverage. In contrast, characteristics such as accruals and operating profit are consistently missing for 30–40% of firms between 2000–2020 because these characteristics require data that can be missing in Compustat for many different reasons.

Missingness is a unique feature of firms. Table 2 reports correlation summary statistics between measures of missing data and several other firm characteristics. In Panel A, columns 1–4, missingness is defined as the fraction of characteristics or input variables that a firm is missing in a given period. In columns 5 and 6, missingness is defined as an indicator variable equal to one if a firm has no annual or quarterly Compustat coverage in a given period, and zero otherwise. In Panels B, C, D, and E, missingness measures are defined as indicator variables equal to one if a firm is missing an individual characteristic (e.g., asset growth) or input variable (e.g., total assets), and zero otherwise.

Results in row 1 of Panel A in Table 2 indicate that the correlation between missing data and firm size is negative but relatively small in magnitude, and ranges from around -8% to around -18%. Results in Panel B indicate that, on an individual characteristic basis, the strongest correlations between missingness and size are around -21%. Thus, while small firms are more likely to have missing financial data relative to large firms, there is not an overly strong relation between firm size and missing data. Figure 2 corroborates these findings, and reports both the average fraction of missing characteristics for firms in each size quintile over time and the fraction of firms in each size quintile with no Compustat coverage over time. Results indicate that, while firms in the largest size quintile consistently have the lowest average missingness, these firms still consistently have a non-trivial fraction of missing Compustat characteristics. In addition, the relation between firm size and missingness has declined over time such that, since around the year 2000, firms in all size quintiles are missing 15-20% of the characteristics on average.

Missing data is negatively correlated with stock return volatility. However, similar to size, the correlations are relatively small in magnitude. Results in Panels A (row 2) and C in Table 2 indicate that the correlation between missingness and return volatility typically ranges from 0% to -18%. These results are generally consistent with Bryzgalova et al. (2023), Chen and McCoy (2023), and Freyberger et al. (2023). Industry membership, exchange membership, and firm age are, in many cases, more highly correlated with missingness than firm size or return volatility. This is because Compustat's data collection and aggregation procedures are related to exchange listing, industry membership, and time since IPO more often than any other factor.

5 Missing Data and Investor Demand

Firms' financial statement information can be important to investors for a wide variety of reasons. Trading strategies can utilize accounting information, investment mandates can incorporate restrictions based on financial ratios, and portfolio managers can use financial data to evaluate the prudence of potential investments. If many investors rely on Compustat to obtain financial statement information, then those investors are implicitly relying on the quality and completeness of Compustat's data coverage. I establish in Section 4 that Compustat's data coverage is not comprehensive. In this Section, I evaluate the connection between missing data in Compustat and institutional investors' equity holdings.

I begin by estimating the following regression:

$$IO_{i,q} = a + b \text{ Missing Data}_{i,q} + cX_{i,q} + FE_q + FE_{SIC2} + FE_{exch} + \epsilon_{i,q} \tag{1}$$

where $IO_{i,q}$ is firm i's level of institutional ownership in quarter q. Missing $\text{Data}_{i,q}$ is an indicator variable defined as 1 if firm i is missing a set of Compustat variables in quarter q. $X_{i,q}$ is a vector of additional firm characteristics (including firm age), FE_q is a time fixed effect, FE_{SIC2} is an industry fixed effect, FE_{exch} is an exchange-listing fixed effect, a is the intercept, and $\epsilon_{i,q}$ is the error term.[14]

5.1 Firms with No Compustat Coverage

In the first set of regressions, I define Missing Data as an indicator variable equal to 1 if a firm has no data available in Compustat for the firm's most recent fiscal year-end. Table 3 reports regression results.

I hypothesize that institutions rely on Compustat to access firms' financial statement information, and that institutions will not invest in firms with missing Compustat data. This implies that missing data in Compustat should affect the extensive margin, or an institution's decision of *whether* to invest. Thus, $FNIO$ is the most relevant measure of institutional ownership. Consistent with this hypothesis, results in Column 2 of Panel A in Table 3 indicate that the fraction of institutional owners is 1.5% lower on average for firms

[14] Despite the fact that I consider many combinations of control variables and fixed effects, endogeneity concerns such as omitted variable bias and reverse causality limit the interpretation of this regression. Section 6 discusses a natural experiment which addresses these concerns.

with no Compustat data coverage. Given that the unconditional mean $FNIO$ is 4%, this is an economically large effect: institutional ownership of firms with no Compustat coverage is over 36% below the mean. This effect is not driven by micro-cap stocks. I find very similar results in Column 7 of Panel A, which excludes the smallest 20% of firms. Results are also similar for the Poisson pseudo-likelihood regressions reported in Columns 4–6. The Poisson model is an important robustness check because it better accounts for the fractional nature and log-normal distribution of the institutional holdings data.

Insofar as the quantity of shares held by institutions is a function of the number of institutional owners, $FSIO$ should also be correlated with missingness. However, conditional on an institution investing in a firm, it is not clear whether missing Compustat data should impact the intensive margin, or an institution's decision of how much to invest. Panel B in Table 3 reports regression results where institutional ownership is defined as $FSIO$. Results in Column 2 indicate that when a firm is not covered in the database, the fraction of shares outstanding held by institutions is approximately 5.7% lower (>16% below the unconditional mean) relative to firms with data coverage.

13f institutions are far from a homogeneous group of investors. Different types of institutions are governed by different regulatory standards and face different investment objectives and mandates. In Panels C and D in Table 3, I report regression results for institutional ownership measures constructed based on the legal type of institution and institutions' size, measured via total equity under management. All missing data coefficient estimates in Panels C and D are significantly negative and economically large in magnitude. Firms with no Compustat coverage have 2.2% (>26% below the unconditional mean) lower bank ownership, 1.3% (>20% below the unconditional mean) lower mutual fund company ownership, and 0.9% (>35% below the unconditional mean) lower investment company ownership relative to firms with data coverage. Results in Panel D show that the negative association between institutional demand and missing data is largest in magnitude for the largest institutions. This effect is likely mechanical: the larger the institution, on average, the more firms they will invest in and therefore the more binding their aversion is to missing data. Results are also robust to inflation-adjusting the institution size reporting threshold, and to weighting institutional ownership measures based on total equity under management.

Missing data in Compustat is most relevant at the institution level because database subscriptions are likely purchased by the institution and made available to all fund managers within the institution. Thus, all fund managers within an institution should face similar data constraints: either the institution subscribes to Compustat or it does not, and either the individual funds are limited by Compustat's data coverage or they are not. For this reason, in the majority of empirical analyses, I focus on investor demand at the institution (fund family) level. However, as a robustness check, I also consider demand at the individual fund level. In these cases, I use the mutual fund holdings data from the Thomson Reuters s12 file. Results in Panel E in Table 3 focus on the association between individual mutual funds' portfolio holdings and missing data. I find that firms with missing Compustat data have approximately 0.33% lower mutual fund ownership (>28% below the unconditional mean). This is is an economically large effect and is consistent with the aggregated results presented throughout Panels A, B, C, and D in Table 3.

Analysts are themselves information intermediaries, and analyst reports are used to inform and advise investors (Healy and Palepu, 2001). It is likely that analysts utilize firms' past financial statement information when developing earnings forecasts and investment recommendations. Similar to institutional investors, if many of the brokerage firms employing analysts and delegating analyst coverage rely primarily on the Compustat database to obtain this information, then these firms are also implicitly relying on the completeness of Compustat's data coverage. I evaluate this possibility in Panel F of Table 3. In this case, I define the left-hand-side variable from eq. (1) as Analyst Coverage$_{i,q}$, which is equal to the number of unique analysts covering firm i in quarter q. Consistent with the notion that the brokerage firms employing analysts rely primarily on the Compustat database to obtain firms' financial statement information, results in Panel F of Table 3 indicate that firms with missing

Compustat information are covered by approximately 2-3 fewer analysts (>44% below the unconditional mean). This is an economically large effect.

5.2 Heterogeneity in Missing Data

Results in Table 3 indicate institutional ownership is over 36% below the unconditional mean for firms with no Compustat coverage. This suggests that whether a firm has any data available in Compustat is a critical determinant of institutional demand. However, there is considerable heterogeneity in the frequency and causes of missingness for many potentially important financial statement variables. For example, Selling, General, and Administrative Expenses (XSGA), Interest Expenses (XINT), and Deferred Taxes (TXDB) are each missing for around 10-20% of firm-fiscal-years *with* Compustat coverage.

I next evaluate the relative importance of missing values for different types of financial statement items. This is an important consideration because, ex ante, the extent to which investors value different types of financial statement information is unclear. The FASB states that the purpose of financial statements is to help market participants accurately forecast firms' future cash flows (Financial Accounting Standards Board, 1978). Presumably, the more granular the accounting information, the more accurate the forecasts and therefore investors should demand very granular financial statement data. However, if investors face relatively binding capacity constraints, then they may focus on only a subset of financial statement items which they deem most relevant and representative of firms' overall operations and performance. Empirically, the literature finds mixed evidence on the relevance of different types of accounting information.[15]

I again estimate regressions as in eq. (1), however, in this case, I define Missing Data as an indicator variable equal to 1 if a firm is missing a specific input variable of interest, and 0 otherwise. Results are presented in Table 4. Because missing data is correlated across various subsets of Compustat items, I begin by focusing on one group of firm characteristics:

[15] Related studies include Healy and Palepu (2001), Holthausen and Watts (2001), Kothari (2001), Kothari et al. (2010), Chen et al. (2015), Leuz and Wysocki (2016), Blankespoor et al. (2020), and Kothari et al. (2023), among many others.

investment, valuation, or profitability. I then break down several popular measures of the associated category of characteristic into their component parts. In cases where two or more input variables are missing only if Compustat does not cover the firm, I include only one Missing Data indicator variable. As an example, asset growth, investment growth, investment-to-capital, and capex growth are four common measures of investment. These characteristics are constructed using combinations of total assets, capital expenditures, total inventory, and total property, plant, and equipment (Compustat items AT, CAPX, INVT, and PPENT, respectively). In Panel A of Table 4, I report results from regressions of institutional ownership on Missing Data indicators for AT, CAPX, and INVT. I do not include a missing indicator variable for PPENT because, in nearly all cases, if a firm has any Compustat coverage, both AT and PPENT are available. Panel B reports results for valuation characteristics, and Panel C reports results for profitability characteristics.

Consistent with the notion that institutional investors rely on Compustat to obtain firms' financial statement information, all Missing Data coefficient estimates for variables that indicate whether a firm has any Compustat coverage (total assets, stockholders' equity, and total sales) are negative and both statistically and economically significant. The magnitudes are also very stable across specifications, and indicate that (aggregate) ownership is 1.4-1.6% lower for firms with no coverage in the most recent fiscal year. In contrast, institutional demand is not consistently related to missing values of other, more nuanced accounting items. For example, missing values of capital expenditures, total inventory, and selling, general, and administrative expenses are not significantly associated with aggregate institutional ownership. There a some cases where the correlation between missing values of more granular variables, such as deferred taxes and investment tax credits (TXDITC) and research and development expenditures (XRD), and institutional ownership are significant; however, these results are not robust to the alternative combinations of fixed effects or regressions of changes in ownership on changes in missingness.

Collectively, the results in Table 4 are consistent with the hypothesis that institutional investors face relatively binding capacity constraints. As a result of these constraints, institutions appear to focus primarily on a subset of financial statement items which are broadly representative of firms' overall operations and performance.

5.3 Robustness Checks

I consider many additional robustness checks with respect to the results reported in Tables 3 and 4. These include Fama MacBeth regression specifications, alternative definitions of 'Missing Data' including changes in missingness, and alternative combinations of control variables. Selected results appear in the Online Appendix.

The choice of control variables included in the regressions in Table 3 is important for at least two reasons. First, although previous literature has linked a wide variety firm characteristics to variation in institutions' stock holdings, many firm characteristics are potentially endogenous to institutional ownership. For example, Koijen and Yogo (2019) develop a model in which institutional demand and stock prices are determined jointly in equilibrium. Bennet et al. (2003) suggest that trading volume and institutional ownership are endogenous. Alti and Sulaeman (2012) suggest that firms' decision to issue a secondary equity offering is influenced by institutional demand. Crane et al. (2016) suggest that higher institutional ownership causes firms to increase dividend payouts. Collectively, this suggests that market capitalization and other firm characteristics that incorporate price, trading volume, share issuance, and dividend yield are all plausibly endogenous to institutional ownership.

Second, because I focus explicitly on the association between incomplete Compustat data coverage and institutional demand, it is imperative that firm-quarter observations with missing Compustat data are included in the sample. This is problematic for any control variable that relies on Compustat data: for any Compustat-based control, when Compustat does not provide all relevant inputs, that control variable is missing and a value must be imputed. In the Internet Appendix, I show that several alternative imputation methods lead to distinct kinks in the values of many control variables. Because these kinks occur when a firm transitions from having missing data to non-missing data, they may bias ‘Missing Data’ regression coefficients. For these reasons, in primary specifications, I include only those controls which do not require Compustat data and do not incorporate variables that prior literature has suggested are endogenous to institutional demand. However, I confirm that empirical results are robust to all combinations of controls.

6 Identification

The previous Section establishes that institutional ownership is significantly lower for firms that are not covered in Compustat. Reasonable endogeneity concerns limit the interpretation of these results. For example, results in Section 5 do not preclude the possibility that S&P caters the database to focus on only those firms for which clients demand information – in this case, institutional demand would determine data coverage. This concern is similar to themes highlighted in D’Souza et al. (2010), who emphasize the interdependence between Compustat’s information dissemination speeds and institutional investor demand. Likewise, results in Section 5 may not address all concerns related to correlated omitted variables (such as the underlying salience of the firm) which potentially cause both variation in data coverage and variation in institutional holdings.

I use a quasi-natural experiment to confirm a plausibly causal connection between Compustat data coverage and institutional investor demand. Specifically, there was a policy change instituted at Standard & Poor’s in the 1990s which drastically increased Compustat’s data coverage. Section 6.1 describes the policy change and its impact on the database. Section 6.2 presents results for the associated difference-in-differences analysis. Section 6.3 describes various robustness exercises.

6.1 Background

Compustat’s data coverage was relatively limited when Standard & Poor’s began collecting financial statement information in 1962. In an effort to improve the database, S&P instituted various policy changes over the following decades, which were often related to either 1) expanding the set of financial statement items collected; or 2) expanding the set of covered public firms.[16] One such event creates an ideal setting to evaluate the causal impact of missing data in Compustat on investor demand: prior to 1993, Compustat covered only a small subset of financial services firms. In 1993, S&P instituted a policy change where they expanded coverage on a going-forward basis to include all financial services firms and back-filled data for many of the previously uncovered financial firms. This 1993 change in data coverage was unrelated to any changes in financial firm's 10-K's or 10-Q's, and was instead triggered by an improvement in the technologies used to collect and process data for financial services firms.

S&P has always maintained two internal data collection systems: one for firms in financial services industries, and one for firms in all other industries. S&P regularly refers to these two categories of firms as 'banks' and 'industrials', respectively.[17] S&P maintains a separate system for the financial firms because the disclosures for these firms are structured very differently from 'industrial' firms in other industries. S&P then uses a "balancing model" to convert financial firms' accounting data into the 'industrial' firm format. In the early 1990s, S&P was able to significantly enhance the financial firm internal data collection system. This positive technological shock enabled them to 1) expand coverage on a going-forward basis to include all (non-newly publicly listed) firms in the financial services industry, 2) back-fill data for many previously uncovered financial firms, and 3) update the balancing model so that many variables which were previously missing for most or all financial firms were no longer missing (e.g., cost of goods sold and total inventory).

This database expansion lead to a discrete, precipitous change in missingness in Compustat for financial firms. Figure 3 shows the percentage of financial firms versus non-financial

[16] For example, S&P began comprehensively covering quarterly 10-Q data for NASDAQ firms in 1983. Likewise, S&P added quarterly research and development expenditures (XRDQ) to the Compustat database in 1989.

[17] To be clear, S&P's 'bank' category refers to firms with SIC codes in the 6000s, which includes financial services firms such as insurance companies and brokerage firms. Thus, it is *not* accurate to interpret this designation as referring to only commercial and investment banks, or to only bank holding companies.

firms with missing values of a variety of popular variables over time. In order to focus on the 1993 policy change and avoid variation in coverage related to newly listed firms in the mid-late 1990s, all firms in Figure 3 are required to be publicly listed in or before Q1 1988. Panel A focuses on missing data in the Compustat North America database. There is a clear, abrupt reduction in missingness in 1994 for financial firms: approximately 50% of financial services firms do not have data in the Compustat North America database prior to the 1993 fiscal year-end. In contrast, missingness is consistently very close to 0 for non-financial firms for all of the period 1988-1999.

Panel B of Figure 3 also shows the percentage of financial firms versus non-financial firms with missing data over time, this time focusing on the Compustat Point-In-Time database. The PIT database states *when* data became available in the Compustat North America database. This is an important robustness check because, during the database expansion in 1993 and 1994, Standard & Poor's back-filled financial data for a significant fraction of financial services firms. Thus, from a backward-looking perspective, the Compustat North America database overstates the amount of information that was available to investors in real-time. Results in Panel B of Figure 3 show that nearly 75% of financial services firms (≈1,000 firms) had no data available prior to 1993. The gradual darkening of the 'Financial Firms' figure in Panel B over 1993 and 1994 reflects both the release of back-filled information for some of the previously uncovered firms, and the attribution of 1993 fiscal year-end information for all financial firms.

Panel B of Figure 3 illustrates that current financial statement information was not comprehensively available for all financial services firms until around the end of 1994. This delay occurs for two related reasons. First, the majority of firms have December 31 fiscal year-ends, which means that accounting data from the 1993 fiscal year is not filed with the SEC until early 1994. This policy change stipulated that Compustat would begin comprehensively covering financial firms' financial statements from the 1993 *fiscal year* onward. As such, it is sensible that this data does not become available for most firms in the database until 1994.

Second, because the Compustat database was undergoing a significant expansion during this time, there was a significant production lag in the attribution of accounting data from the 1993 fiscal year into the database. This meant that financial statement information from the 1993 fiscal year-end did not become available in the database for many firms until the third or fourth quarter of 1994. From 1995 onward, nearly 100% of the financial and non-financial firms (excluding new listings) are covered in Compustat and have non-missing values of basic financial statement items such as income, assets, and stockholder's equity.

6.2 Difference-in-Differences Analysis

Standard & Poor's 1990's policy change, and its associated increase in Compustat's coverage of financial services firms, was driven by a positive shock to S&P's data collection technologies. It therefore provides a setting to evaluate the causal connection between Compustat data coverage and institutional investor demand. If institutional investors rely on Compustat to access firms' financial statement information then, following the positive shock to coverage, ownership should increase.

I use a difference-in-differences approach to test the causal relation between missing data in Compustat and investor demand:

$$IO_{i,q} = a + b(\text{Treated}_i \times \text{Post}_q) + c\ \text{Treated}_i + dX_{i,q} + FE_q + FE_{SIC2} + FE_{exch} + \epsilon_{i,q} \quad (2)$$

where $IO_{i,q}$ is firm i's level of institutional ownership in quarter q. I define treated firms as all financial firms with no data in the Compustat Point-In-Time database prior to 1993; Treated is an indicator that equals one for these firms, and zero otherwise. I follow Standard & Poor's definition of financial services, and classify all firms with SIC codes ranging from 6000–6999, excluding codes 6411, 6792, 6794, and 6795, as financials. Because the treatment affect is dispersed across 1993 and 1994, and is not complete until the final quarter of 1994, Post is defined as an indicator that equals one in and after 1995, and zero otherwise. The regression sample spans the period Q1 1988 – Q4 1999, and includes only those firms which publicly listed in or before Q1 1988.

Regression results are reported in Table 5. The b coefficient estimate for $\text{Treated}_i \times \text{Post}_q$ is consistently significant and positive across all specifications. This indicates that, following the change in coverage, institutional ownership increased by around 1% for the subset of treated firms. Notably, results are robust to time, industry, exchange, and firm fixed effects, to the inclusion of time-varying controls, and to alternative functional forms.

Panel A of Table 5 reports results for aggregate institutional ownership. Panel B reports results for legal types of institutions. Panel C reports results for various institution sizes. Panel D reports results for individual mutual fund holdings and for analyst coverage. Consistent with the discussion in Section 5, ownership levels increase for all types and sizes of institutions following the initiation of Compustat coverage. Results are also similar for analyst coverage and alternative measures of mutual fund ownership.

Figure 4 evaluates the parallel trends requirement. Panel A reports the cross-sectional average institutional ownership for treated (dotted-blue line) versus untreated (solid orange line) firms from Q1 1988 through Q4 1999. The left-hand figure reports the average FNIO, and the right-hand figure reports the average NIO, demeaned by firms' NIO in Q1 1988. There are two shaded regions in the figures. The light-grey, diagonal slash shaded region indicates the period over which back-filled data for (some) treated firms began to appear in Compustat. The darker-grey, horizontal slash shaded region indicates the period over which treated firms' most recent fiscal-year end data began to appear in Compustat.

The left-hand chart in Panel A in Figure 4 shows that treated and untreated firms have very similar average trends in institutional ownership from 1988 – early 1993, but that treated firms have nearly 1.5% lower ownership (equivalent to 35–40% smaller magnitude) relative to untreated firms. Around the end of 1993, the average level of ownership for treated firms begins to gradually approach the average level of ownership for untreated firms. By the end of 1997, the average level of ownership is the same for treated and untreated firms. The gradual convergence of treated firms' average level of institutional ownership to untreated firms' average level of ownership is exactly what is expected under the hypothesis that institutions avoid investing in firms with missing data. If many institutional investors rely on Compustat to obtain firms' financial statement data, then these investors would not hold stock in firms with missing data. Once a firm begins to be covered in Compustat, these investors *might* invest in the firm; however, the decision of whether or not to invest likely depends on many other factors, such as the value of various financial ratios.

Results in the right-hand chart in Panel A in Figure 4 illustrate that this convergence in average FNIO for treated versus control firms is driven by accelerated growth in the number of institutional owners of treated firms throughout the mid-to-late 1990s; the number of institutional owners for control firms is also increasing over this period, but it is increasing at a slower rate. Panel B in Figure 4 reports dynamic treatment effects over time, and illustrates that the b coefficient estimate from regression eq. (2) becomes large, positive, and statistically significant following the positive shock to Compustat coverage.

Altogether, results in Figure 4 and Table 5 are consistent with the conclusion that (many) institutional investors rely on Compustat to access firms' financial statement information, and that they will not invest in firms that are not covered in the database.

6.3 Robustness Checks and Alternative Explanations

I consider a variety of robustness checks with respect to the empirical results presented in Section 6.2. These include several placebo and falsification tests such as varying the definition of 'Post' and randomly assigning the treatment effect across firms. Explicit results appear in the Online Appendix, as do parallel trends figures evaluating an extended sample period and additional summary statistics for the treated and control samples. The Online Appendix also presents results for samples which drop firms that engage in a merger or acquisition during the evaluation period, and for samples where treated and control firms are matched based on firms' size and level of institutional ownership, both measured in Q1 1988. Regression results for these alternative samples are very similar to those presented in Table 5.

This quasi-natural experiment treats the majority financial services firms over the period 1993-1994. While regression results in Table 5 are robust to the inclusion of both industry and firm fixed effects, this does not rule out the possibility that an alternative event occurred around the same time frame that 1) affected only financial services firms, and 2) caused the increase in institutional ownership for these firms. I explore alternative explanations in the following subsections.

6.3.1 The Rise in Sector Specific Investing

The number of sector and industry specific mutual funds and exchange-traded funds grew substantially during the 1990s and 2000s. It is possible that, in their development of these funds, investment and mutual fund companies realized that there was significant under-investment in the financial services industry and created more financial-services-oriented funds to fill this gap.

It is reasonable to conjecture that this rise in sector specific investing could explain the increase in institutional ownership for financial firms, relative to all other firms, during the mid-1990s. However, empirical evidence in Panel B of Table 5 is inconsistent with this hypothesis. Under this alternative explanation, the increase in aggregate institutional ownership should be driven by mutual fund companies and investment companies/advisors. Results in Panel B illustrate that all types of institutions, including banks, insurance companies, and pension funds, increased their ownership of treated firms following the treatment effect. Analysts also increased their coverage of treated firms following the treatment effect. This is inconsistent with the notion that the rise in sector specific investing drove the change in ownership for financial services firms following the increase in Compustat's coverage of these firms.

6.3.2 Deregulation of the Banking Industry

The Riegle-Neal Interstate Banking and Branching Efficiency Act was passed in 1994. This act removed the restrictions which previously prevented banks from engaging in interstate banking and from branching across state lines. Literature examining the impact of this regulation has largely concluded that it increased the competitiveness of U.S. banking markets (Zarutskie, 2006; Rice and Strahan, 2010). Ex ante, it is plausible that this affected institutional demand for banks between 1993 and 1997.

The Riegle-Neal Act did not affect all banks equally because states maintained the authority to create barriers to branch expansion. Specifically, states could limit interstate branching in any of the following four ways: First, states could limit interstate bank mergers by setting a minimum age requirement for all target institutions. Second, states could cap the percentage of deposits controlled by any single bank or bank holding company, thus limiting banks' ability to engage in large interstate mergers. Third, de novo interstate branching was only permitted if states decided to "opt-in" to this feature of the regulation. Finally, interstate mergers of individual branches was also only permitted if states decided to "opt-in" to this feature of the regulation. Collectively, this means that interstate branching was only possible via whole-bank mergers which met minimum age requirements and did not exceed the relevant deposit cap for states that elected not to opt-in to these provisions. Rice and Strahan (2010) exploit variation in states' adoption of these different barriers to entry to create a state-level index of branching restrictiveness.

Under the hypothesis that the passage of the Riegle-Neal Act explains the increase in institutional ownership for financial firms in the mid-1990s, the increase in institutional ownership should be largest for firms located in states with the most open branching laws post-Riegle-Neal. This is because banks in more open states were more affected by Riegle-Neal than banks in less open states. In Panel E of Table 5, I examine whether there is any variation in the treatment effect across firms located in different states. In contrast to this hypothesis, I find no evidence that variation in branching restrictions is related to institutional demand: the increase in institutional ownership for treated firms relative to control firms is significantly positive and consistent in magnitude for firms located in both very open and very restricted states. This is inconsistent with the notion that the Riegle-

Neal Act drove the change in ownership for financial services firms following the increase in Compustat's coverage of these firms.

6.3.3 EDGAR

The SEC's EDGAR database was implemented on a staggered schedule between 1993–1996. Studies such as Gao and Huang (2020) and Kim et al. (2023) highlight the role that EDGAR played in massively reducing information acquisition costs. While it is likely that many institutional investors began using EDGAR as soon as it became available, the introduction of EDGAR alone can not explain the differential trends in institutional investment for treated financial services firms versus all other firms in the 1990s.

If institutional investors primarily relied on SEC resources to obtain firms' financial statement information in the 1990s, then the introduction of EDGAR should have substantially reduced the costs they incurred to collect that information. However, under this hypothesis, there is little reason to believe that EDGAR's introduction should have differentially affected institutions' decision to invest in financial services firms relative to firms in other industries. Alternatively, if institutional investors relied primarily on Compustat to obtain firms' financial statement information in the 1990s, then both the increase in Compustat's coverage of financial firms and EDGAR's implementation would plausibly contribute to the differential increase in ownership for financial firms during the mid-1990s. The empirical results presented in Section 6.2, combined with existing evidence indicating that EDGAR is not most investors' primary source of public disclosures (Pricewaterhouse Coopers, LLP, 2006, June 8; Drake et al., 2015; Blankespoor et al., 2020), supports the conclusion that the change in Compustat's data coverage in the 1990s drove the increase in institutional investment of financial services firms.

7 Who Uses Compustat?

The empirical analyses in Sections 5 and 6 show that, in aggregate, a significant fraction of institutional investors do not invest in firms with no Compustat data coverage. This confirms Compustat's relevance as an information intermediary. It also highlights a surprising quality of the professional asset management industry: even though it is possible to supplement Compustat with self-collected data, the empirical results suggest that many institutions do not do so. In this Section, I evaluate several potential explanations for why this is the case.

7.1 Hypothesis Development

First and foremost, institutional investors manage assets on clients' behalf. As noted in studies such as Lakonishok et al. (1992) and Edelen et al. (2022), there are many agency conflicts that arise because of this principal-agent relation. One way in which the asset management industry mitigates these agency costs is via investment mandates, which broadly specify funds' investment strategies and investable assets. Insofar as these investment mandates incorporate and/or allow for the use of accounting information, they should also affect institutions' reliance on Compustat data. For example, a smart beta fund might consider accounting-based firm characteristics such as asset growth, operating profitability, and the book-to-market ratio when making investment decisions, and it may rely explicitly on Compustat to obtain this accounting data. In contrast, a passive index fund invests only in index constituents and therefore its portfolio allocations should not be directly related to Compustat's coverage. Likewise, funds engaging in high turnover strategies based on high frequency information (e.g., price) may require little or no accounting information. In comparison, funds engaging in strategies which incorporate lower frequency accounting information will be directly affected by Compustat's coverage if the fund utilizes the database.

There are two empirical predictions that arise from this first 'investment mandate' hypothesis. First, passive index funds should be more inclined to invest in firms with no

Compustat data relative to actively managed funds. Second, the very highest turnover funds, which are the most likely to engage in strategies focusing on past returns and other high frequency information, should be more inclined to invest in firms with no Compustat data relative to lower- and mid-turnover funds.

An additional way in which agency costs in the asset management industry are mitigated is via regulatory constraints. As fiduciaries, institutional investors have a legal obligation of due diligence. While a wide variety of firm characteristics could reasonably be used to defend an investment decision, both the academic literature and the legal case history routinely cite fundamental ratios, which incorporate accounting information, as evidence of the prudence of an investment.[18] It is possible that these due diligence requirements deter portfolio managers from self-collecting data. For example, portfolio managers may require data for various financial ratios so that they can clearly communicate to stakeholders why they made certain investment decisions, and why those investments satisfy the relevant prudence requirements. They may be reluctant to self-collect data because this adds the additional burden of defending the integrity and comparability of their self-collected data, relative to data obtained from a well-established data vendor. Studies such as Harris and Morsfiled (2012) highlight the complexities of the taxonomy underlying firms' financial statements, and emphasize that standardizing accounting information in a meaningful way across firms, industries, and over time is very much a non-trivial task.

This 'due diligence' hypothesis suggests that those institutions that are relatively more constrained by agency conflicts should be relatively more averse to self-collecting data. Thus, institutions facing more stringent regulatory environments should be less inclined to self-collect data relative to institutions facing more lenient regulatory environments. Likewise, younger and smaller institutions with weaker reputations should be less inclined to self-collect data relative to older, more established institutions.

[18] An example of a court case where fundamental, accounting-based characteristics were used to evaluate the prudence of an investment decision is *First Alabama Bank of Montgomery, N.A. v. Martin (1983)*. Examples of academic studies which evaluate the prudence of an investment using similar characteristics include Badrinath et al. (1989, 1996); Del Guercio (1996); Falkenstein (1996)

A third potential explanation for why some institutions may rely more exclusively on a data vendor than others is the explicit cost of data collection. From an equilibrium perspective, it is only optimal for portfolio managers to self-collect data if the marginal benefits of obtaining the information are greater than the marginal costs associated with collecting that information (Grossman and Stiglitz, 1980). It is possible that, for many institutions, the explicit costs associated with acquiring firms' disclosures and maintaining a database of information for these uncovered firms exceed the benefits they might accrue from obtaining and trading on the additional accounting data. Studies such as Gao and Huang (2020), Bowles et al. (2023), and Kim et al. (2023) highlight many issues associated with obtaining 10-K and 10-Q disclosures from the SEC, and emphasize that this is *not* a cost-less endeavor.

This 'explicit cost' hypothesis predicts that those investors that benefit relatively more from acquiring the public disclosures should be relatively more inclined to self-collect data. This suggests that more actively managed funds and larger funds should be more inclined to self-collect data relative to less actively managed funds and smaller funds. This is because funds that engage in active stock picking are designed to generate positive alpha by identifying individual firms which are mispriced, and necessarily have more discretion in determining portfolio allocations across assets. In contrast, less actively managed funds, such as those based on factor beta or smart beta strategies, involved little (if any) stock picking and are instead designed to accrue returns based on risk exposure with relatively more fixed portfolio allocations. To the extent that activeness reflects the portfolio manager's skill, the same prediction arises: the more skilled the manager, the better their ability to identify mispricing, and therefore the more likely they are to incur the cost of self-collecting data. Likewise, high fixed costs associated with data collection and economies of scale suggest that smaller institutions are more likely to find it optimal to rely exclusively on a data vendor, while larger institutions are more likely to augment Compustat with self-collected data.

7.2 Empirical Analyses

Section 7.1 describes three alternative explanations for why institutional investors may rely exclusively on Compustat, and why some institutions may chose not to self-collect data for firms not covered in the database. These hypotheses are not mutual exclusively and, in fact, it is possible that all three play a role in influencing institutions' incentive to rely exclusively on a data vendor. In order to evaluate these hypotheses, I examine institutional investors' propensity to invest in firms with missing Compustat data.

I estimate the following regression at the institution-quarter level:

$$\begin{aligned}\%\text{ Portfolio No Data}_{j,q} = a + b\text{ Log(EUM)}_{j,q} + c\text{ Age}_{j,q} + d\text{ Turnover}_{j,q} + \\ \text{Legal Type FE}_j + \text{Time FE}_q + \epsilon_{j,q} \quad (3)\end{aligned}$$

where % Portfolio No Data$_{j,q}$ is the fraction of institution j's portfolio invested in firms with no Compustat data in quarter q. Log(EUM)$_{j,q}$ is the log of institution j's total equity under management, Age is the number of quarters that the institution has appeared in the 13f database, and Turnover is the institution's portfolio turnover defined as in Yan and Zhang (2009). Legal Type FE$_j$ is a set of fixed effects reflecting institutions' legal types. I consider two alternative definitions of '% Portfolio No Data': the fraction of the institution's equity under management invested in firms with no Compustat data ('Fraction of EUM') and the fraction of firms that the institution holds with no Compustat data ('Fraction of Firms').

Throughout the majority of this study, I focus on institution-level data because database subscriptions are likely maintained at the institution (or fund family) level. However, investment mandates, which often govern investment strategies, benchmark indices, and activeness, are typically specified at the fund level. For this reason, I also estimate similar regressions using the s12 data regarding individual mutual funds:

$$\begin{aligned}\%\text{ Portfolio No Data}_{f,q} = a + b\text{ Log(EUM)}_{f,q} + c\text{ Age}_{f,q} + d\text{ Turnover}_{f,q} + e\text{ Active Share}_{f,q} \\ + \text{ Index Fund}_{f} + \text{Enhanced Index Fund}_{f} + \text{Time FE}_{q} + \epsilon_{f,q} \quad (4)\end{aligned}$$

where, in this case, % Portfolio No Data$_{f,q}$ is the fraction of mutual fund f's portfolio invested in firms with no Compustat data in quarter q. Active Share is defined as in Cremers and Petajisto (2009) and Petajisto (2013), and is equal to the percentage of a fund's portfolio holdings that differ from the fund's benchmark index. Data regarding funds' active share are obtained from Annti Petajisto's website.[19] Table 6 reports regression results. Panel A focuses on institutional investors and regressions as in eq. (3), while Panel B focuses on individual mutual funds and regressions as in eq. (4).

The 'investment mandate' hypothesis predicts that 1) passive index funds are more likely to invest in firms with missing data relative to non-index funds, and 2) the highest turnover funds are more likely to invest in firms with missing data relative to lower- and mid-turnover funds. Consistent with the first prediction, results in Panel B of Table 6 show that index funds and enhanced index funds invest a significantly larger fraction of their portfolios in firms with missing Compustat data compared to non-index funds. Consistent with the second prediction, results in Panels A and B of Table 6 show that Portfolio Turnover is positively associated with the fraction of an institution's and mutual fund's portfolio invested in firms with missing data. For example, institutions in the lower three turnover quintiles each have approximately 0.25-0.27% less of their equity under management invested firms with no Compustat data coverage relative to institutions in the highest turnover quintile. This is equivalent to approximately 8% of a standard deviation difference in 'Fraction of EUM.'

The 'due diligence' hypothesis predicts that institutions facing more stringent regulatory environments should be less inclined to self-collect data relative to institutions facing more lenient regulatory environments. Studies such as Badrinath et al. (1989), Badrinath et al. (1996), and Del Guercio (1996) emphasize that insurance companies, banks, and pension

[19] https://www.petajisto.net/data.html. This data is available only for domestic, all equity mutual funds, which are not sector funds and which have a minimum of $10 million in assets under management.

funds are subject to much stricter prudent-man laws compared to investment companies and advisors. Consistent with the 'due diligence' hypothesis, results in Panel A of Table 6 show that insurance companies, banks, and pension funds invest a significantly smaller fraction of their portfolios in firms with missing Compustat data compared to mutual fund and investment companies.

The 'due diligence' hypothesis also predicts that younger and smaller institutions with weaker reputations should be more constrained by Compustat's data coverage relative to older, more established institutions. Likewise, the 'explicit cost' hypothesis predicts that total assets under management should positively predict an institutions' propensity to invest in firms with missing Compustat data. Support for these size- and age-related predictions is limited. In Panel A, the institution age coefficient estimate varies in sign and significance, and the institution size coefficient is generally insignificant. In Panel B, the mutual fund age coefficient is insignificant in all but one regression, while the mutual fund size coefficient is significantly negative only when mutual fund fixed effects are included. This inconsistency in empirical results may be arise because 1) institution and mutual fund age and equity under management are imperfect indicators of individual portfolio managers' reputational capital, and 2) the empirical EUM distribution is dominated by a small number of extremely large institutions/funds, and these funds tend to be the least active.

The 'explicit cost' hypothesis predicts that activeness should be positively associated with a fund's propensity to invest in firms with missing data. Consistent with this, results in Panel B of Table 6 show that Active Share is positively related to the fraction of a mutual fund's portfolio invested in firms with missing data. Mutual funds in the highest active share quintile (the 'active stock pickers') have 1.1% more of their equity under management invested firms with no Compustat data coverage relative to mutual funds in the lowest active share quintile (the 'closest indexers'). This is equivalent to over 50% of a standard deviation difference in 'Fraction of EUM.'

Collectively, the results in Table 6 provide support for all three hypotheses. This suggests that investment mandates, due diligence constraints, and the explicit costs associated with self-collecting data all influence institutional investors' incentive to rely on Compustat as an information intermediary.

7.3 Fund Performance

Do institutional investors and portfolio managers who invest in firms with missing Compustat data perform better than their counterparts, who appear to rely more exclusively on the data vendor? To the extent that an institutions' propensity to invest in firms with missing Compustat data reflects their ability to utilize self-collected data and/or their access to a superior data source (e.g., an alternative data vendor with superior coverage), those institutions that rely more exclusively on Compustat should under-perform their less-constrained counterparts. I explore this possibility in the Online Appendix. I find that there is a positive correlation between an institution's or mutual fund's propensity to invest in firms with missing Compustat data and their future performance. However, the correlation is not consistently statistically significant and is dominated by other characteristics, such as activeness. This is consistent with the conclusion that more skilled and less constrained institutions tend to both 1) have access to and/or collect superior accounting data, and 2) tend to perform better than less skilled and more constrained institutions.

8 Missing Data and Information Assimilation

If financial statement data is informative, and if investors rely on Compustat to obtain this data, then the absence of Compustat coverage should limit information assimilation in financial markets. In Sections 5 and 6, I show that a significant fraction of institutions do not invest in firms with missing data in Compustat, confirming the database's relevance as an information intermediary. I next evaluate the connection between Compustat's data coverage and the informational efficiency of equity prices.

There are two potential channels through which Compustat coverage may influence information assimilation in financial markets. First, it is possible that financial statement data informs investors' forecasts of and reactions to news releases. Under this hypothesis, investors' ability to precisely forecast and/or accurately react to information releases may be limited when they do not have access to prior accounting data. For example, if financial statement information helps market participants forecast firms' future cash flows, then in the absence of this information, investors' earnings forecasts will be less accurate and therefore they are more likely to be surprised by the information contained in an earnings announcement. Second, it is possible that firms not covered in Compustat face substantially less scrutiny by many market participants who require accounting information (e.g., for due diligence reasons) before investing. Under this hypothesis, these uncovered firms' equity prices will be significantly less informationally efficient because of limited investor attention (in the spirit of 'neglected' firms discussed in Merton (1987)), and, while the past financial statement data may be informative regarding the prudence of an investment, it need not be directly valuation-relevant.

I consider several alternative settings to evaluate the connection between Compustat coverage and information assimilation. In each setting, I construct firm-level empirical proxies for stock price informational inefficiency ('II'), and estimate the following regression:

$$\begin{aligned} II_{i,t} = a + b\text{Missing Data}_{i,t-1} + c\left(\text{Missing Data}_{i,t-1} \times \text{Investor Attention}_{i,t-1}\right) \\ + d\text{ Investor Attention}_{i,t-1} + eX_{i,t-1} + FE_t + FE_{SIC2} + FE_{exch} + \epsilon_{i,t} \quad (5) \end{aligned}$$

where $II_{i,t}$ is the relevant informational inefficiency measure for firm i at time t, Missing $\text{Data}_{i,t-1}$ is an indicator variable defined as 1 if firm i is not covered in Compustat in the $t-1$ fiscal year, and Investor $\text{Attention}_{i,t-1}$ is a proxy for market participation, measured as either institutional ownership or analyst coverage. I hypothesize that missing Compustat data mitigates information assimilation, and that this effect will be partially offset by increased investor attention – that is, that $b > 0$ and $c < 0$ in regression (5).

8.1 Quarterly Earnings Announcements

I begin by evaluating the connection between Compustat data coverage and returns during and after quarterly earnings announcements. I use quarterly earnings announcements as a laboratory from which to study information assimilation because 1) these announcements provide firm-specific, valuation-relevant, fundamental information at definitive points in time, and 2) it is well-documented in the empirical literature that there is a significant price drift following these announcements (e.g., Ball and Brown, 1968; Fink, 2020).

I follow prior literature (Blankespoor et al., 2020; Fink, 2020) and estimate announcement period cumulative abnormal returns ('CARs') over the window $\tau = [-1, 1]$, where $\tau = 0$ is the earnings announcement date. I define post-announcement CARs over the window $\tau = [2, 60]$. In main results, I focus on two alternative models for the normal return: the single-factor market model and the Fama and French (1993) three-factor model.[20] I consider alternative windows and alternative normal return models as robustness checks. Finally, I estimate regressions as in eq. (5), where the dependent variable is defined as the announcement period or post-announcement absolute cumulative abnormal return, $ACAR_{i,\tau}$. Earnings announcements are measured in each quarter of year t, the 'Missing Data' indicator reflects whether a firm has any Compustat data coverage for the $t-1$ fiscal year-end, and all other variables are measured as of the end of the quarter prior to the earnings announcement.

Table 7 reports regression results. Panel A focuses on earnings surprises. Results in columns 1–4 indicate that announcement period returns are 0.3–0.5% larger in magnitude for firms with no Compustat coverage for the most recent fiscal year-end. Results also indicate that this effect is offset if there are sufficiently many analysts covering the firm and/or sufficiently many institutions investing in the firm. Specifically, the interaction coefficient estimates suggest that an increase of approximately 6-8 ($\approx\geqslant$1 standard deviation increase) analysts or an increase in institutional ownership of 5-15% ($\approx\geqslant$1 standard deviation increase)

[20] Under the single-factor market model, I define all $\beta_{mkt,i} = 1$. This first approach avoids issues associated with estimating betas. Under the Fama and French (1993) three-factor model, I use a 250-trading-day window ending on day $\tau - 2$ to estimate factor loadings.

negates the impact of missing Compustat data.

Panel B of Table 7 focuses on post-announcement drift. Results in columns 1–4 indicate that post-announcement returns are 0.7–1% larger in magnitude for firms with no Compustat coverage for the most recent fiscal year-end. Results also indicate that this effect is offset if there are sufficiently many analysts covering the firm and/or sufficiently many institutional investors, however the effect is both statically weaker and smaller in magnitude than that for earnings surprises. Specifically, the interaction coefficient estimates suggest that an increase of at least 7 analysts or an increase in institutional ownership of at least 9% negates the impact of missing Compustat data.

The results in Table 7 support the conclusion that limited access to financial statement data limits the informational efficiency of equity prices: when Compustat does not cover a firm, earnings surprises are larger, post-earnings announcement drift is larger, and information assimilation is slower. These effects are mitigated if investor attention is sufficiently high. This suggests that Compustat data coverage affects information assimilation in financial markets via its impact of market participation and investor attention.[21]

8.2 Return Autocorrelations and Price Delay

I next consider the connection between Compustat data coverage and more general proxies for information assimilation, including measures of daily return autocorrelations and measures of price delay.

French and Roll (1986) argue that the absolute levels of firms' daily return autocorrelations should be positively related to investors' mis-reactions to new, firm-specific information. Thus, a firm's autocorrelation coefficient (ρ) serves as a measure for the informational inefficiency of the firm's stock price: in an informationally efficient market, prices reflect all public information and returns should follow a random walk (i.e., $\rho \approx 0$). For this reason, I evaluate

[21] It is important to note that I do not directly observe institutional investors' or analysts' information sets. It is therefore difficult to definitively rule out the possibility that the offsetting effect of increased attention arises because these investors and analysts have access to an alternative source of financial statement data (such as self-collected data), which causes them to be better informed and to improve information assimilation.

the connection between Compustat data coverage and firms' return autocorrelations. I first estimate the following AR(1) regression:

$$R_{i,d} = \alpha_i + \rho_i R_{i,d-1} + \epsilon_{i,d} \tag{6}$$

where $R_{i,d}$ is the daily return for firm i on trading day d. I estimate these regressions at the firm-level using one year of daily returns, requiring a minimum of 100 trading days of data.

In addition to measures of daily return autocorrelations, I consider the three alternative measures of 'price delay' proposed by Hou and Moskowitz (2005), which are designed to estimate the delay with which firms' stock prices incorporate market-wide information. These measures are obtained from the following regressions:

$$R_{i,w} = \alpha_i + \beta_i^0 R_{MKT,w} + \sum_{n=1}^{4} \left(\beta_i^{-n} R_{MKT,w-n}\right) + \epsilon_{i,w} \tag{7}$$

where $R_{i,w}$ is the weekly return for firm i in week w, and $R_{MKT,w}$ is the value-weighted market return in week w. Weekly returns are measured from Wednesday–Tuesday. I estimate these regressions at the firm-level using one year of weekly returns, requiring a minimum of 24 weeks of data.

The three measures of price delay proposed in Hou and Moskowitz (2005) are then computed as follows. The first is equal to the fraction of variation in a firm's returns explained by lagged market returns:

$$D1 = 1 - \frac{R^2_{\beta_i^{-n}=0,\forall n \in [1,4]}}{R^2} \tag{8}$$

where R^2 is the r-squared from regression (7), and $R^2_{\beta_i^{-n}=0,\forall n \in [1,4]}$ is the r-squared from regression (7) when restricting $\beta_i^{-n} = 0$ for $\forall n \in [1,4]$. Intuitively, the larger the value of $D1$, the more return variation is captured by lagged market returns and therefore the stronger firm i's delay is in response to market-wide return innovations.

The second and third measures of price delay are designed to distinguish between shorter and longer lags, and to account for the precision of the β estimates:

$$D2 = \frac{\sum_{n=1}^{4} n\beta^{-n}}{\beta^0 + \sum_{n=1}^{4} \beta^{-n}} \tag{9}$$

$$D3 = \frac{\sum_{n=1}^{4} (n\beta^{-n}/se(\beta^{-n}))}{(\beta^0/se(\beta^0)) + \sum_{n=1}^{4} (\beta^{-n}/se(\beta^{-n}))} \tag{10}$$

where $se(\beta^{-n})$ is the standard error of the relevant coefficient estimate. The intuition behind these alternative measures is straightforward: if stock i's price responds immediately to market news, then β_i^0 will be significantly different from zero, while none of the β_i^{-n}'s will differ from zero. However, if stock i's price responds with a lag, then some or all of the β_i^{-n}'s will differ significantly from zero.

In order to evaluate the connection between Compustat data coverage and these alternative measures of information assimilation, I estimate regressions as in eq. (5) at the annual frequency. In daily return autocorrelation regressions, II is defined as $|\rho|$, $|\frac{\rho}{se(\rho)}|$, or the r-squared from regression (6). In price delay regressions, II is defined as $D1$, $D2$, or $D3$. In all cases, the informational inefficiency measures are constructed using stock return data from July in year t through June in year $t+1$. The 'Missing Data' indicator reflects whether a firm has any Compustat data coverage for the $t-1$ fiscal year-end, and all other variables are measured as of the end of June in year t.

Results are reported in Table 8 and uniformly indicate that missing Compustat data is associated with stronger return autocorrelations and stronger price delays. Results in Panel A indicate that the autocorrelation coefficients are over 0.02 larger in magnitude (≈20% of a standard deviation) for firms with no Compustat data relative to firms with Compustat coverage. Similarly, the autocorrelation t-statistics more than 0.3 higher (≈20% of a standard deviation), and the r-squared values from the AR(1) regressions are 1.5% higher (≈50% of a standard deviation), for firms with missing data. Likewise, results in Panel B indicate that the fraction of stock-specific return variance captured by lagged market returns is approximately 4.6% larger (≈15% of a standard deviation) for firms with no Compustat data relative to firms with Compustat coverage. Results are again similar across alternative price delay measures: the ratio of lagged β^{-n} coefficients to the sum of all β coefficients is approximately 0.2 higher (≈7% of a standard deviation) for firms with missing data. The similar ratio for β t-statistics is 0.16 higher (≈5% of a standard deviation) for firms with missing data.

Collectively, results in both Panels of Table 8 are consistent with the notion that equity prices are less informationally efficient for firms that are not covered in Compustat: when Compustat does not cover a firm, return autocorrelations are larger, price delay measures are larger, and information assimilation is slower. Results in Panel B are also consistent with the conclusion that this effect is offset if there are sufficiently many analysts covering the firm and/or sufficiently many institutions investing in the firm. Although the interaction coefficient estimates in Panel A are uniformly insignificant, the interaction coefficients in Panel B suggest that an increase of approximately 6-10 analysts or an increase in institutional ownership of 6-10% offsets the impact of missing Compustat data on price delay measures. These results are collectively consistent with the earnings announcement analysis in Section 8.1, and suggest that limited access to financial statement data reduces the informational efficiency of equity prices via its impact on market participation and investor attention.

9 Conclusion

Many theoretical studies highlight costs associated with information acquisition and processing as a first-order friction in financial markets (Grossman and Stiglitz, 1980; Merton, 1987). However, empirically measuring the marginal costs and benefits of becoming informed is challenging. In this paper, I propose a novel empirical setting that tackles this challenge. I use this setting to directly evaluate the influence that information availability can have on investor demand, and how limited access to information affects market efficiency.

There is an entire industry of professional data aggregators who collect and standardize data on clients' behalf. These data vendors act as information intermediaries in a wide variety of contexts. Standard & Poor's Compustat database is one of the oldest and most prominent data vendors in the finance industry, and Compustat has contributed to massive reductions in the collective cost of aggregating and processing public firms' accounting information. However, the database is not and has never been comprehensive. I examine how the completeness of Compustat's data coverage affects institutional investor demand. I then examine how Compustat coverage affects information assimilation in equity markets.

I first show that missing data is a pervasive and non-trivial problem in Compustat, and that institutional ownership of firms without Compustat coverage is over 36% below its unconditional mean. I use a quasi-natural experiment to confirm a plausibly causal connection between Compustat data coverage and institutional demand. I then evaluate the connection between Compustat coverage and information assimilation. Consistent with the conclusion that limited access to accounting information reduces the informational efficiency of equity markets, in a battery of empirical tests, I find that stock prices of firms with missing Compustat data are significantly less informationally efficient relative to firms with more complete data coverage.

This study highlights the role that data vendors play in facilitating the flow of information within the economy and emphasizes frictions that arise in the context of this information intermediation. Although many of the frictions related to the intermediation of financial statement information have attenuated in recent years, financial statement data is only one subset of potentially relevant information, and Compustat is only one data vendor. Recent decades witnessed continuous and exponential improvements in information technologies and data gathering methods. Data vendors such as Glassdoor, the Carbon Disclosure Project, the Privacy Rights Clearinghouse, StockTwits, and other social media platforms now provide information related to employee satisfaction, pollution, cybersecurity risk, retail investor sentiment, and other ESG-related firm characteristics. All of this information is plausibly relevant to a significant fraction of investors. As such, the role that data vendors play as information intermediaries will continue to be relevant in studies of financial markets.

References

Akbas, F., Markov, S., Subasi, M., Weisbrod, E., 2018. Determinants and consequences of information processing delay: Evidence from the thomson reuters institutional brokers' estimate system. Journal of Financial Economics 127, 366–388.

Alti, A., Sulaeman, J., 2012. When do high stock returns trigger equity issues? Journal of Financial Economics 103, 61–87.

Badrinath, S., Gay, G. D., Kale, J. R., 1989. Patterns of institutional investment, prudence, and the managerial "safety-net" hypothesis. Journal of Risk and Insurance 56, 605–629.

Badrinath, S., Kale, J. R., Ryan, H. E., 1996. Characteristics of common stock holdings of insurance companies. Journal of Risk and Insurance 63, 49–76.

Badrinath, S., Wahal, S., 2002. Momentum trading by institutions. Journal of Financial Markets 57, 2449–2478.

Ball, R., Brown, P., 1968. An empirical evaluation of accounting income numbers. Journal of Accounting Research 6, 159–178.

Barringer, F., Fabrikant, G., March 21, 1999. Coming of age at bloomberg l.p. New York Times, Available at: https://www.nytimes.com/1999/03/21/business/coming-of-age-at-bloomberg-lp.html?searchResultPosition=1 p. 1.

Ben-Rephael, A., Da, Z., Israelsen, R. D., 2017. It depends on where you search: Institutional investor attention and underreaction to news. Review of Financial Studies 30, 3009–3047.

Bennet, J., Sias, R., Starks, L., 2003. Greener pastures and the impact of dynamic institutional preferences. Review of Financial Studies 16, 1203—1238.

Bird, A., Karolyi, S. A., 2016. Do institutional investors demand public disclosure? Review of Financial Studies 29, 3245–3277.

Blankespoor, E., deHaan, E., Marinovic, I., 2020. Disclosure processing costs, investors' information choice, and equity market outcomes: A review. Journal of Accounting and Economics 70, 101344.

Boritz, J. E., No, W. G., 2020. How significant are the differences in financial data provided by key data sources? a comparison of xbrl, compustat, yahoo! finance, and google finance. Journal of Information Systems 34, 47–75.

Bowles, B., Reed, A. V., 2024. Mutual fund shorts and the benefits of acquiring information. Working Paper .

Bowles, B., Reed, A. V., Ringgenberg, M. C., Thornock, J. R., 2023. Anomaly time. Journal of Finance, forthcoming .

Bryzgalova, S., Lerner, S., Lettau, M., Pelger, M., 2023. Missing financial data. Working Paper .

Bushee, B. J., Matsumoto, D. A., Miller, G. S., 2003. Open versus closed conference calls: The determinants and effects of broadening access to disclosure. Journal of Accounting and Economics 34, 149–180.

Bushee, B. J., Noe, C. F., 2000. Corporate disclosure practices, institutional investors, and stock return volatility. Journal of Accounting Research 38, 171–202.

Chen, A. Y., McCoy, J., 2023. Missing values and the dimensionality of expected returns. Working Paper .

Chen, S., Miao, B., Shevlin, T., 2015. A new measure of disclosure quality: The level of disaggregation of accounting data in annual reports. Journal of Accounting Research 53, 1017–1054.

Chu, Y., Hirshleifer, D., Ma, L., 2020. The causal effect of limits to arbitrage on asset pricing anomalies. Journal of Finance 75, 2631–2672.

Chuk, E., Matsumoto, D., Miller, G. S., 2013. Assessing methods of identifying management forecasts: Cig vs. researcher collected. Journal of Accounting and Economics 55, 23–42.

Chychyla, R., Kogan, A., 2015. Using xbrl to conduct a large-scale study of discrepancies between the accounting numbers in compustat and sec 10-k filings. Journal of Information Systems 29, 37–72.

Correia, S., Guimarães, P., Zylkin, T., 2019. Verifying the existence of maximum likelihood estimates for generalized linear models.

Correia, S., Guimarães, P., Zylkin, T., 2020. Fast Poisson estimation with high-dimensional fixed effects. The Stata Journal 20, 95–115.

Crane, A. D., Crotty, K., Umar, T., 2023. Hedge funds and public information acquisition. Management Science 69, 3241–3262.

Crane, A. D., Michenaud, S., Weston, J. P., 2016. The effect of institutional ownership on payout policy: Evidence from index thresholds. Review of Financial Studies 29, 1377–1408.

Cremers, K. M., Petajisto, A., 2009. How active is your fund manager? a new measure that predicts performance. Review of Financial Studies 22, 3329–3365.

Da, Z., Engelberg, J., Gao, P., 2011. In search of attention. Journal of Finance 66, 1461–1499.

Dasgupta, A., Pratt, A., Verardo, M., 2011. Institutional trade persistence and long-term equity returns. Journal of Finance 66, 635–653.

Del Guercio, D., 1996. The distorting effect of the prudent-man laws on institutional equity investment. Journal of Financial Economics 40, 31–62.

Detzel, A., Novy-Marx, R., Velihov, M., 2023. Model comparison with transaction costs. Journal of Finance, forthcoming .

Drake, M. S., Roulstone, D. T., Thornock, J. R., 2015. The determinants and consequences of information acquisition via edgar. Contemporary Accounting Research 32, 1128–1161.

Drechsler, Q. F. S., 2023. Python programs for empirical finance. Available at: https://www.fredasongdrechsler.com .

D'Souza, J., Ramesh, K., Shen, M., 2010. The interdependence between institutional ownership and information dissemination by data aggregators. The Accounting Review 85, 159—-193.

Du, K., Huddart, S., Jiang, X., 2023. Lost in standardization: Effects of financial statement database discrepancies on inference. Journal of Accounting and Economics 75.

Easterwood, S., 2024. Why is data missing in crsp and compustat? Working Paper .

Edelen, R. M., Hosseinian, A., Kadlec, G. B., 2022. The investable universe of 13f institutions. Working Paper .

Edelen, R. M., Ince, O., Kadlec, G. B., 2016. Institutional investors and stock return anomalies. Journal of Financial Economics 119, 472–488.

Falkenstein, E., 1996. Preferences for stock characteristics as revealed by mutual fund portfolio holdings. Journal of Finance 51, 111–135.

Fama, E. F., French, K. R., 1992. The cross-section of expected returns. Journal of Finance, 46, 427–466.

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

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

Farboodi, M., Matray, A., Veldkamp, L., Venkateswaran, V., 2022. Where has all the data gone? Review of Financial Studies 35, 3101–3138.

Financial Accounting Standards Board, 1978. Statement of financial accounting concepts no.1: objectives of financial reporting by business enterprises. FASB, Stamford, CT .

Fink, J., 2020. A review of the post-earnings-announcement drift. Journal of Behavioral and Experimental Finance 29, 1–13.

Frazzini, A., Israel, R., Moskowitz, T. J., 2014. Trading costs of asset pricing anomalies. Working Paper .

French, K. R., Roll, R., 1986. Stock return variances: The arrival of information and the reaction of traders. Journal of Financial Economics 17, 5–26.

Freyberger, J., Hoppner, B., Neuhierl, A., Weber, M., 2023. Missing data in asset pricing panels. Working Paper .

Gao, M., Huang, J., 2020. Informing the market: The effect of modern information technologies on information production. Review of Financial Studies 33, 1367–1411.

Goldstein, I., Yang, L., 2017. Information disclosure in financial markets. Annual Review of Financial Economics 9, 101–125.

Gompers, P., Metrick, A., 2001. Institutional investors and equity prices. Quarterly Journal of Economics 116, 229–259.

Grossman, S. J., Stiglitz, J. E., 1980. On the impossibility of informationally efficient markets. American Economic Review 70, 393–408.

Gutierrez, R., Kelley, E., 2009. Institutional herding and future stock returns. Working Paper .

Harris, T., Morsfiled, S., 2012. An evaluation of the current state and future of xbrl and interactive data for investors and analysts. Columbia Business School Center for Excellence in Accounting and Security Analysis, White Paper Number Three.

Hasbrouck, J., 1991. Measuring the information content of stock trades. Journal of Finance 46, 179–207.

Healy, P., Palepu, K., 2001. Information asymmetry, corporate disclosure, and the capital markets: A review of the empirical disclosure literature. Journal of Accounting and Economics, 31, 405–440.

Hoitash, R., Hoitash, U., 2018. Measuring accounting reporting complexity with xbrl. The Accounting Review 93, 259–287.

Holthausen, R. W., Watts, R. L., 2001. The relevance of the value-relevance literature for financial accounting standard setting. Journal of Accounting and Economics 31, 3–75.

Hong, H., Lim, T., Stein, J. C., 2000. Bad news travels slowly: Size, analyst coverage, and the profitability of momentum strategies. Journal of Finance 55, 265–295.

Hou, K., Moskowitz, T. J., 2005. Market frictions, price delay, and the cross-section of expected returns. Review of Financial Studies, 18, 981–1020.

Johnston, J. A., Reichelt, K. J., Sapkota, P., 2024. Measuring financial statement disaggregation using xbrl. Journal of Information Systems 38, 119–147.

Jones, C. M., Lamont, O. A., 2002. Short-sale constraints and stock returns. Journal of Financial Economics 66, 207–239.

Karpoff, J. M., Koester, A., Lee, D. S., Martin, G., 2017. Proxies and databases in financial misconduct research. The Accounting Review 92, 129–163.

Kim, Y. H., Ivkoich, Z., Muravyev, D., 2023. Causal effect of information costs on asset pricing anomalies. Working Paper .

Koijen, R., Yogo, M., 2019. A demand system approach to asset pricing. Journal of Political Economy 127, 1475–1515.

Kothari, S., 2001. Capital markets research in accounting. Journal of Accounting and Economics 31, 105–231.

Kothari, S., Ramanna, K., Skinner, D., 2010. Implications for gaap from an analysis of positive research in accounting. Journal of Accounting and Economics 50, 246–286.

Kothari, S., Zhang, L., Zuo, L., 2023. Disclosure regulation: Past, present, and future. Handbook of Financial Decision Making, forthcoming .

Lakonishok, J., Shleifer, A., Vishny, R. W., 1992. The structure and performance of the money management industry. Brookings Papers on Economic Activity 23, 339–391.

Lesmond, D. A., Schill, M. J., Zhou, C., 2004. The illusory nature of momentum profits. Journal of Financial Economics 71, 349–380.

Leuz, C., Wysocki, P., 2016. The economics of disclosure and financial reporting regulation: Evidence and suggestions for future research. Journal of Accounting Research 54, 525–622.

Lewellen, J., 2011. Institutional investors and the limits of arbitrage. Journal of Finance 102, 62–80.

Ljungqvist, A., Malloy, C., Marston, F., 2009. Rewriting history. Journal of Finance 64, 1935—1960.

Love, J. P., 1992, January 14. Public access to sec information. IU Bia-Archive http://www.bio.net/bionet/mm/ag-forst/1992-January/000187.html .

Merton, R. C., 1987. A simple model of capital market equilibrium with incomplete information. Journal of Finance 42, 483–510.

Nagel, S., 2005. Short sales, institutional investors and the cross-section of stock returns. Journal of Financial Economics 78, 277–309.

Noble, K. B., 1982. Sec data: Difficult hunt. The New York Times .

Novy-Marx, R., Velikov, M., 2016. A taxonomy of anomalies and their trading costs. Review of Financial Studies 29, 104–147.

Petajisto, A., 2013. Active share and mutual fund performance. Financial Analysts Journal 69, 73–93.

Pricewaterhouse Coopers, LLP, 2006, June 8. Sec letter. Available at: https://www.sec.gov/news/press/4-515/4515-8.pdf .

Rice, T., Strahan, P. E., 2010. Does credit competition affect small-firm finance? Journal of Finance 65, 861–889.

Schaub, N., 2018. The role of data providers as information intermediaries. Journal of Financial and Quantitative Analysis 53, 1805–1838.

Shleifer, A., Vishny, R., 1997. The limits of arbitrage. Journal of Finance 52, 35–55.

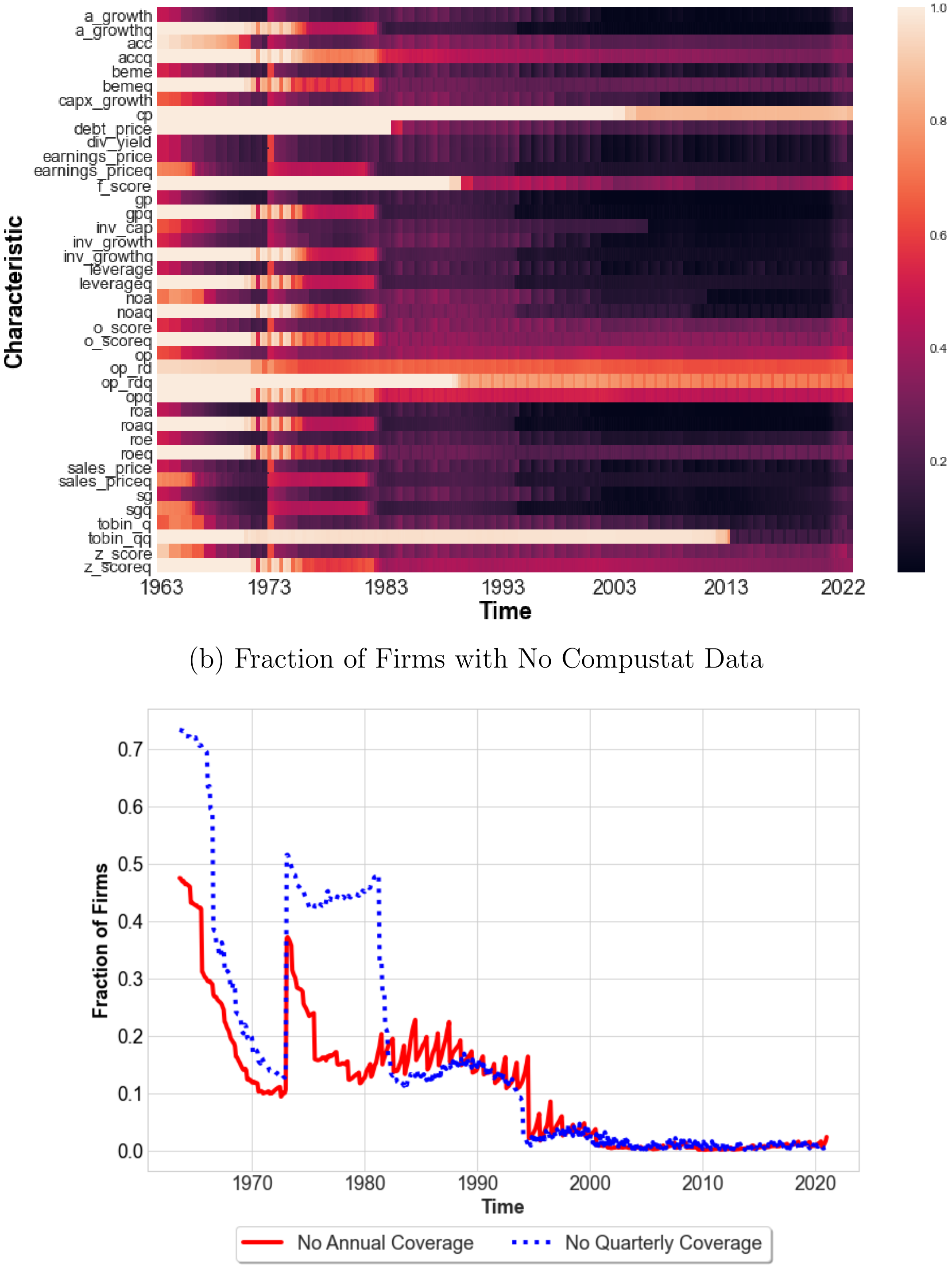
Yan, X., Zhang, Z., 2009. Institutional investors and equity returns: Are short-term institutions better informed. Review of Financial Studies 22, 893–924.

Zarutskie, R., 2006. Evidence on the effects of bank competition on firm borrowing and investment. Journal of Financial Economics 81, 503–537.

Figure 1: **Missing Characteristic Data Over Time**

This figure reports the fraction of firms with missing Compustat data over time. Panel A shows the fraction of firm-month observations with missing values for a variety of accounting-based characteristics. Panel B shows the fraction of firm-month observations with no Compustat coverage over time. Characteristics, abbreviations, and definitions are reported in Appendix Table 1.

(a) Fraction of Firms with Missing Characteristic Data

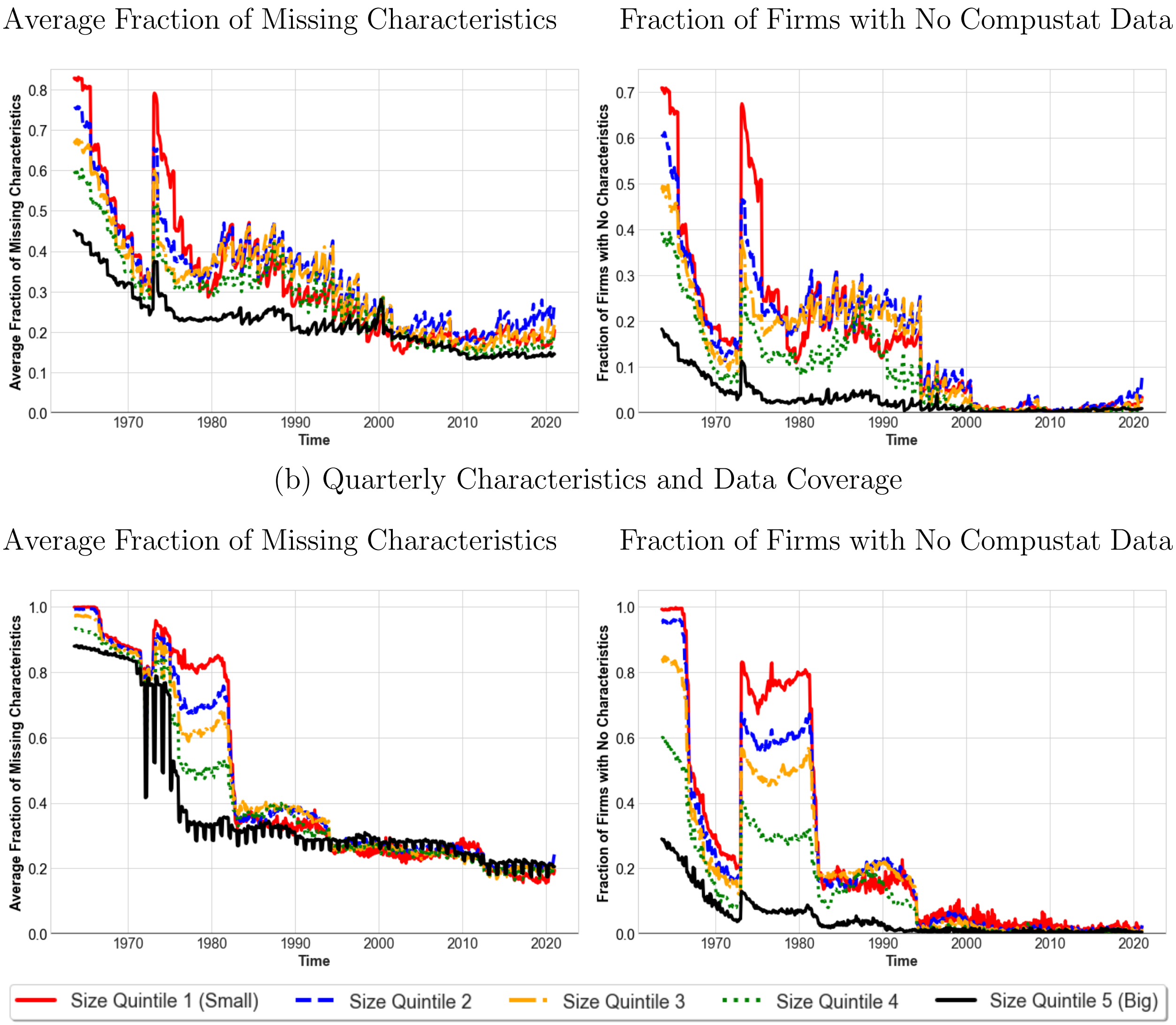


(b) Fraction of Firms with No Compustat Data

Figure 2: **Missing Characteristic Data Over Time By Firm Size**

This figure displays missing data summary statistics for firms in each of five size quintiles over the time. Left-hand figures report the average fraction of missing characteristics for firms in each size quintile over the time. Right-hand figures report the fraction of firms with no Compustat data in each size quintile over the time. Characteristics correspond to the annual (Panel A) and quarterly (Panel B) characteristics that also appear in Figure 1. Size quintiles are defined based on market capitalization measured as of the end of the prior month.

(a) Annual Characteristics and Data Coverage



Average Fraction of Missing Characteristics

Fraction of Firms with No Compustat Data

(b) Quarterly Characteristics and Data Coverage

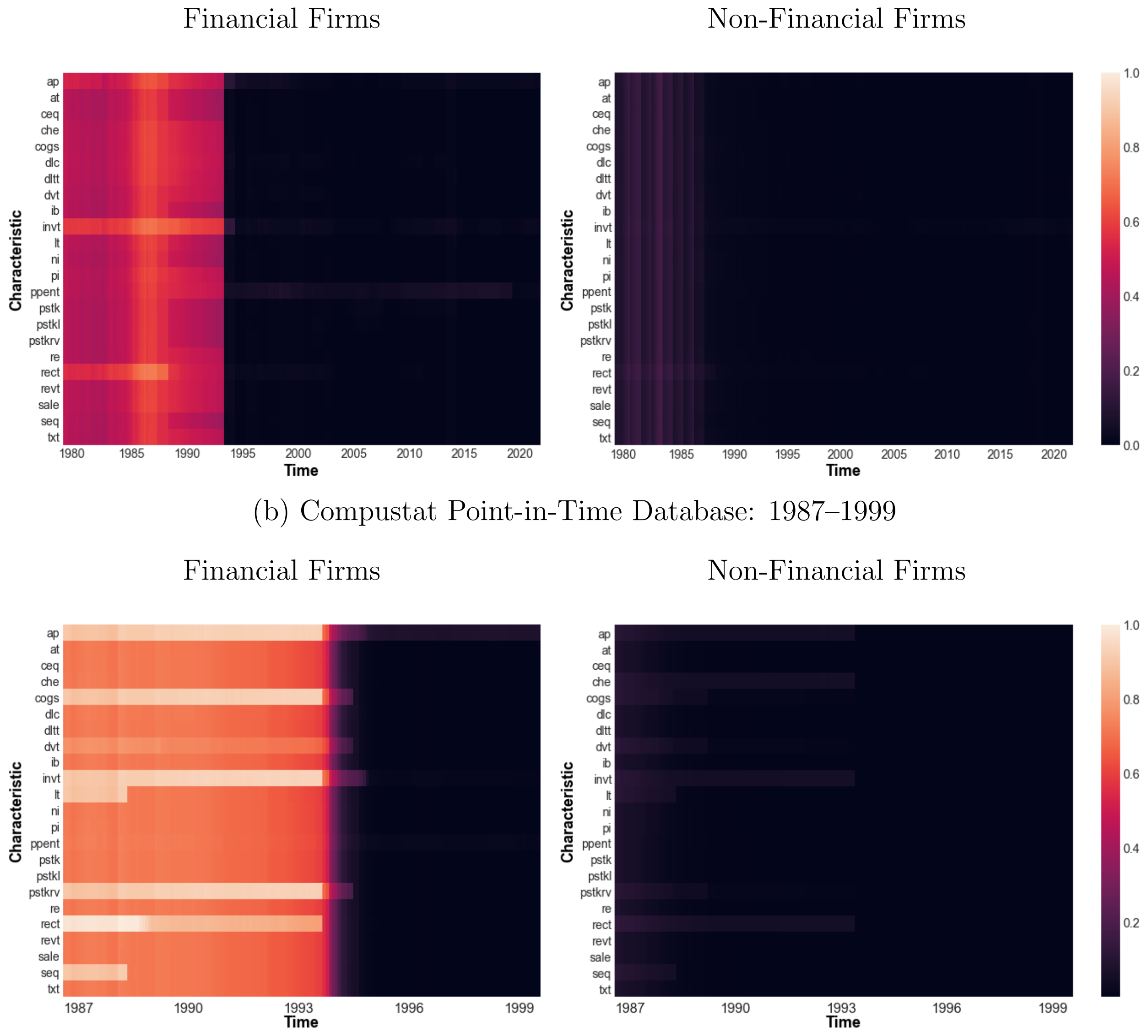
Average Fraction of Missing Characteristics

Fraction of Firms with No Compustat Data

Figure 3: **Change in Compustat Data Coverage for Financial Firms**

This figure displays the percentage of firm-months with missing values for input variables obtained from Compustat which are used to construct a variety of popular firm characteristics. Panel A defines data as missing if it is missing in the standard Compustat North America database. Panel B defines data as missing if it was not available in Compustat in real-time; this is measured using the Compustat Point-in-Time database. Left-hand figures show results for financial services firms. Standard & Poor's classifies firms with SIC codes ranging from 6000 – 6999, excluding codes 6411, 6792, 6794, 6795, as financial services. Right-hand figures show results for all other firms. All firms are required to be publicly listed in or before Q1 1988.

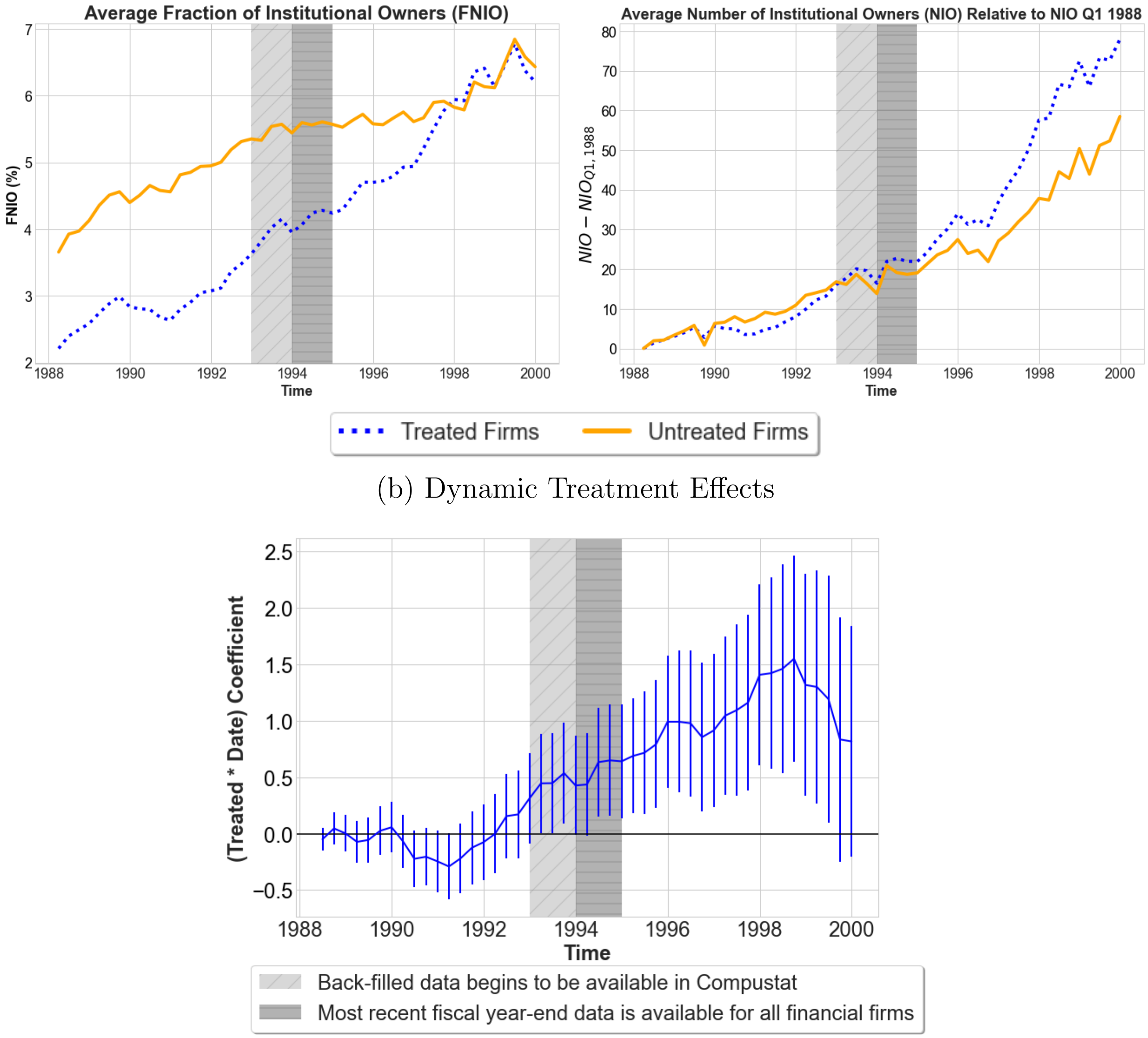
(a) Compustat North America Database: 1980–2021



(b) Compustat Point-in-Time Database: 1987–1999

Figure 4: **Parallel Trends**

This figure displays results for various tests of parallel trends. Panel A reports the cross-sectional average institutional ownership for treated versus untreated firms between Q1 1988 and Q4 1999. The left-hand figure reports average FNIO for treated (dotted-blue line) versus untreated (solid-yellow line) firms. The right-hand figure reports average NIO relative to each firm's NIO in Q1 1988. Panel B reports Treated × Date coefficient estimates from regression (2), including a 99% confidence interval. Treated firms are defined as all financial services firms with no data in the Compustat Point-in-Time database prior to 1993. Compustat classifies firms with SIC codes ranging from 6000 – 6999, excluding codes 6411, 6792, 6794, 6795, as financial services. Untreated firms are defined as all other firms. Both treated and untreated firms are required to have market capitalization data in CRSP in Q1 1988. The light-grey, diagonal slash shaded region indicates the period over which back-filled data for (some) treated firms began to appear in the Compustat North America database. The darker-grey, horizontal slash shaded region indicates the period over which treated firms' most recent fiscal-year end data began to appear in the Compustat North America database.



(a) Average Institutional Ownership

(b) Dynamic Treatment Effects

Table 1: **Summary Statistics**

This table presents summary statistics for Institutional Ownership, Mutual Fund Ownership, and Analyst Coverage. $FNIO_{i,q}$ is equal to the number of institutions that hold shares of stock i in quarter q, scaled by the total number of institutions in the 13f data set (s34 file) in quarter q. $FSIO$ is equal to the fraction a stock's shares outstanding held by institutions. $FNMF_{i,q}$ is equal to the number of mutual funds that hold shares of stock i in quarter q, scaled by the total number of mutual funds in the 13f data set (s12 file) in quarter q. Analyst Coverage is equal to the number of unique analysts covering a stock.

Panel A: FNIO (%)

	Mean	Std	10%	Median	90%	Fraction = 0
Full Sample	4.03	6.96	0.12	1.58	10.23	6.6
1980s	3.61	7.84	0.0	0.79	9.90	17.8
1990s	3.73	6.83	0.11	1.32	9.64	5.4
2000s	4.39	6.42	0.26	2.31	10.50	0.9
2010s	4.56	6.48	0.27	2.64	10.88	1.0

Panel B: FSIO (%)

	Mean	Std	10%	Median	90%	Fraction = 0
Full Sample	34.77	30.96	0.39	026.53	83.78	6.6
1980s	16.11	18.42	0.0	8.82	45.17	17.8
1990s	27.14	24.17	0.58	20.81	64.26	5.4
2000s	44.50	31.38	3.64	42.28	88.71	0.9
2010s	57.95	32.85	5.68	66.50	96.34	1.0

Panel C: FNIO by Institution Type (%)

Institution Type:	Insurance Company	Bank	Pension Fund	Mutual Fund Company	Investment Company/ Advisor	Miscellaneous
Mean	8.16	8.17	9.37	6.40	2.48	2.23
Std	12.00	13.53	14.79	9.18	4.94	4.89
10%	0	0	0	0	0	0
Median	2.56	3.18	2.17	2.43	0.87	0
90%	23.94	20.75	30.19	18.06	6.14	6.45

Panel D: Fraction of Institutions w/in Type Classifications (%)

Institution Type:	Insurance Company	Bank	Pension Fund	Mutual Fund Company	Investment Company/ Advisor	Miscellaneous
Full Sample	1.99	7.74	2.15	5.14	72.49	10.49
1980s	7.74	27.43	6.17	15.88	37.20	5.58
1990s	4.09	15.47	3.17	17.46	56.07	3.74
2000s	1.64	5.53	2.15	8.61	72.13	9.94
2010s	1.01	3.30	1.54	4.30	80.92	8.93

Panel E: Mutual Fund Ownership (FNMF, %)

	Mean	Std	10%	Median	90%	Fraction = 0
Full Sample	1.18	2.36	0	0.35	3.01	18.80
1980s	0.89	2.11	0	0.19	2.42	40.60
1990s	0.90	2.05	0	0.21	2.36	20.80
2000s	1.22	2.46	0.02	0.43	3.00	4.60
2010s	1.91	2.76	0.03	1.07	5.43	5.20

Panel F: Analyst Coverage (#)

	Mean	Std	10%	Median	90%	Fraction = 0
Full Sample	4.52	6.39	0	2	14	36.0
1980s	3.35	6.11	0	0	12	53.5
1990s	3.97	6.10	0	1	12	38.2
2000s	4.61	5.81	0	2	13	29.9
2010s	6.67	7.18	0	4	17	18.0

Table 2: **Firm Characteristic Correlations**

This table reports correlations between firm characteristics and various measures of missingness. Firm size is equal to the firm's market capitalization percentile as of the end of the prior month. Return volatility is equal to the firm's return standard deviation percentile as of the end of the prior quarter. Financial Firm is an indicator variable equal to one if a firm has an SIC code in the 6000s, and zero otherwise. NASDAQ is an indicator variable equal to one if a firm is listed on the NASDAQ exchange, and zero otherwise. Age is defined as the number of months since the firm's IPO. In Panel A, columns 1–4, missingness is equal to the fraction of missing characteristics or input variables. 'Firm Characteristics' measures are based on the 40 characteristics in Figure 1. 'Compustat Input Variables' measures are based on the input variables required to construct each of these 40 characteristics. In Panel A, columns 5–6, 'Compustat Coverage' is an indicator variable defined as one if a firm has no data available in Compustat in a given fiscal period, and zero otherwise. In Panels B, C, D, E, and F, missingness is constructed for each individual firm characteristic and is defined as an indicator variable equal to one if the firm characteristic is missing, and zero otherwise. Summary statistics are reported for the correlations between each characteristic's missing data indicator variable and firm size, volatility, financial firm indicator, NASDAQ indicator, and age. The sample period is July 1963 – Dec 2022 for all results except those related to the NASDAQ indicator. Results corresponding to the NASDAQ indicator are estimated over the sample Jan 1973 – Dec 2022.

Panel A: Correlations Between Missingness and Other Firm Characteristics

	Firm Characteristics:		Compustat Input Variables:		Compustat Coverage:	
Missingness Measure:	Annual Characteristics	Quarterly Characteristics	Annual Inputs	Quarterly Inputs	No Annual Coverage	No Quarterly Coverage
Size	-0.1145	-0.0809	-0.1087	-0.1257	-0.1379	-0.1825
Return Volatility	-0.0745	-0.0856	-0.0888	-0.0592	-0.0079	0.0068
Financial Firm Indicator	0.4190	0.2874	0.3804	0.2465	0.2374	0.1717
NASDAQ Indicator	0.1979	0.1452	0.1684	0.1898	0.1751	0.2247
Age (months)	-0.3198	-0.2590	-0.2428	-0.2614	-0.2181	-0.2353

Panel B: Missingness and Firm Size Correlation Summary Statistics

	Min	10%	Median	Avg	90%	Max
Annual Characteristics	-0.1825 (sg)	-0.1493	-0.1013	-0.0953	-0.0299	-0.0047 (op)
Quarterly Characteristics	-0.2154 (sgq)	-0.1468	-0.0755	-0.0743	0.0068	0.0336 (o_scoreq)
Annual Input Variables	-0.1398 (ib)	-0.1364	-0.1128	-0.0913	-0.0206	-0.0010 (xsga)
Quarterly Input Variables	-0.1878 (niq)	-0.1657	-0.1094	-0.0999	-0.0126	0.0165 (xrdq)

Table 2: **Firm Characteristic Correlations**

Panel C: Missingness and Return Volatility Correlation Summary Statistics

	Min	10%	Median	Avg	90%	Max
Annual Characteristics	-0.1527 (acc)	-0.1297	-0.0224	-0.0449	0.0117	0.0415 (sg)
Quarterly Characteristics	-0.1844 (o_scoreq)	-0.1382	-0.0315	-0.0526	0.0119	0.0497 (tobin_qq)
Annual Input Variables	-0.1858 (dvpa)	-0.1742	-0.0278	-0.0625	-0.0164	0.0255 (tstkp)
Quarterly Input Variables	-0.1765 (xrdq)	-0.1492	-0.0139	-0.0427	0.0043	0.0640 (txdbq)

Panel D: Missingness and Financial Firm Indicator Correlation Summary Statistics

	Min	10%	Median	Avg	90%	Max
Annual Characteristics	0.0873 (cp)	0.1520	0.2374	0.2868	0.5301	0.5845 (o_score)
Quarterly Characteristics	0.0137 (tobin_qq)	0.0765	0.1514	0.2021	0.3878	0.4337 (o_scoreq)
Annual Input Variables	0.0074 (xpp)	0.1615	0.2434	0.2891	0.5182	0.7175 (act)
Quarterly Input Variables	0.0183 (txdbq)	0.0869	0.1234	0.1841	0.3778	0.5442 (actq)

Panel E: Missingness and NASDAQ Indicator Correlation Summary Statistics

	Min	10%	Median	Avg	90%	Max
Annual Characteristics	0.0186 (cp)	0.0285	0.1740	0.1532	0.2049	0.2247 (sg)
Quarterly Characteristics	-0.0629 (op_rdq)	0.0400	0.1263	0.1166	0.1850	0.2376 (sgq)
Annual Input Variables	-0.0326 (xrd)	0.0414	0.1693	0.1389	0.1761	0.1942 (txp)
Quarterly Input Variables	-0.1340 (xrdq)	0.0401	0.1524	0.1480	0.2248	0.2890 (xintq)

Panel F: Missingness and Age Correlation Summary Statistics

	Min	10%	Median	Avg	90%	Max
Annual Characteristics	-0.3297 (f_score)	-0.3164	-0.2348	-0.2390	-0.1450	-0.0889 (op_rd)
Quarterly Characteristics	-0.2629 (tobin_qq)	-0.2577	-0.2316	-0.1992	-0.1312	-0.0558 (op_rdq)
Annual Input Variables	-0.2322 (oancf)	-0.2218	-0.2170	-0.1929	-0.1407	-0.0618 (xrd)
Quarterly Input Variables	-0.2573 (txdbq)	-0.2368	-0.2252	-0.2052	-0.1445	-0.0186 (xrdq)

Table 3: **Investor Demand and Missing Data**

This table reports regressions of investor demand, primarily institutional ownership, on an indicator for missing Compustat data, as in eq. (1). ‘Missing Data’ is an indicator variable defined as 1 if a firm has no Compustat data available for its most recent fiscal year-end, and 0 otherwise. In Panels A and B, institutional ownership (FNIO or FSIO) is measured for all institutions in aggregate. In Panels C and D, institutional ownership (FNIO) is measured for individual types and sizes of institutions. ‘Inflation-Adj Size Cutoff’ measures FNIO with respect to only those institutions whose total EUM falls above the inflation-adjusted 13f reporting threshold. (The threshold is equal to $100 million in Q1 1980, and is adjusted over time for inflation.) ‘EUM-weighted’ measures FNIO, where each institution is weighted by their total EUM in the relevant quarter. In Panel E, investor demand is measured via mutual fund ownership (FNMF, s12 file). In Panel F, demand is measure via analyst coverage, which is a count variable equal to the number of unique analysts covering the firm. All measures of demand (except analyst coverage) are expressed in percentage points. Standard errors are clustered by firm and date, t-statistics are reported in parentheses, and marginal effects for non-linear regressions are reported in brackets. Linear regressions are estimated via Ordinary Least Squares. Poisson pseudo-likelihood regressions are estimated as in Correia et al. (2019) and Correia et al. (2020). All regressions are linear unless indicated otherwise. Continuous control variables are winsorized at the 1% and 99% levels. Controls include: the Koijen and Yogo (2019) market cap instrument, an S&P500 indicator, age, and age squared. Industry is defined as 2-digit SIC code. The sample period is 1980–2021.

Panel A: Aggregate Institutional Ownership with Dependent Variable FNIO

	All Firms						Drop Smallest 20%	
Missing Data	-3.2925 *** (-31.09)	-1.4729 *** (-24.45)	-0.3439 *** (-8.54)	-1.4950 *** (-33.85) [-6.0225]	-0.3560 *** (-10.50) [-1.4343]	-0.1696 *** (-7.08) [-0.6919]	1.5054 *** (-20.76)	-0.3934 *** (-8.24)
Regression	Linear	Linear	Linear	Poisson	Poisson	Poisson	Linear	Linear
Controls	N	Y	Y	N	Y	Y	N	Y
Time FE	Y	Y	Y	Y	Y	Y	Y	Y
Industry FE	N	Y	N	N	Y	N	N	Y
Exchange FE	N	Y	N	N	Y	N	N	Y
Firm FE	N	N	Y	N	N	Y	N	N
Adjusted R^2	2.13%	66.11%	90.19%	3.86%	63.08%	69.36%	66.25%	89.91%
N	848,838	848,838	848,499	848,838	848,834	837,947	678,999	678,360

Panel B: Aggregate Institutional Ownership with Dependent Variable FSIO

	All Firms						Drop Smallest 20%	
Missing Data	-12.0925 *** (-36.10)	-5.6754 *** (-18.53)	-2.4382 *** (-9.57)	-0.6456 *** (-21.58) [-22.4436]	-0.3136 *** (-12.43) [-10.9021]	-0.1787 *** (-14.32) [-6.2922]	-5.0561 *** (-14.46)	-2.3527 *** (-8.18)
Regression	Linear	Linear	Linear	Poisson	Poisson	Poisson	Linear	Linear
Controls	N	Y	Y	N	Y	Y	N	Y
Time FE	Y	Y	Y	Y	Y	Y	Y	Y
Industry FE	N	Y	N	N	Y	N	N	Y
Exchange FE	N	Y	N	N	Y	N	N	Y
Firm FE	N	N	Y	N	N	Y	N	N
Adjusted R^2	29.02%	58.15%	83.23%	23.67%	51.80%	72.26%	60.08%	83.03%
N	848,838	848,838	848,499	848,838	848,834	837,947	678,999	678,360

Table 3: **Investor Demand and Missing Data**

Panel C: Institution Legal Type

Institution Type:	Insurance Company	Bank	Pension Fund	Mutual Fund Company	Investment Company/Advisor	Miscellaneous
Missing Data	-1.1041 ***	-2.1734 ***	-1.0672 ***	-1.2958 ***	-0.8793 ***	-0.5063 ***
	(-11.50)	(-20.65)	(-7.52)	(-18.29)	(-17.95)	(-11.40)
Time FE	Y	Y	Y	Y	Y	Y
Industry FE	Y	Y	Y	Y	Y	Y
Exchange FE	Y	Y	Y	Y	Y	Y
Controls	Y	Y	Y	Y	Y	Y
Adjusted R^2	71.84%	69.37%	74.65%	66.39%	56.61%	52.77%
N	848,838	848,838	848,838	848,838	848,838	848,838

Panel D: Institution Size

EUM Quintile:	Q1 (Small)	Q2	Q3	Q4	Q5 (Large)	30 Largest	Inflation-Adj Size Cutoff	EUM- Weighted
Missing Data	-0.5216 ***	-0.7647 ***	-0.9921 ***	-1.4906 ***	-3.6030 ***	-8.1773 ***	-1.7158 ***	-5.5237 ***
	(-12.62)	(-13.65)	(-15.92)	(-20.11)	(-29.98)	(-26.54)	(-25.74)	(-23.70)
Time FE	Y	Y	Y	Y	Y	Y	Y	Y
Industry FE	Y	Y	Y	Y	Y	Y	Y	Y
Exchange FE	Y	Y	Y	Y	Y	Y	Y	Y
Controls	Y	Y	Y	Y	Y	Y	Y	Y
Adjusted R^2	39.91%	45.81%	52.06%	62.60%	75.80%	74.67%	71.28%	77.92%
N	848,838	848,838	848,838	848,838	848,838	848,838	848,838	848,838

Panel E: Mutual Fund Ownership

	All Firms						Drop Smallest 20%	
Missing Data	-0.8692 ***	-0.3329 ***	-0.1037 ***	-1.5460 ***	-0.2649 ***	-0.0504 *	-0.3293 ***	-0.1077 ***
	(-26.24)	(-17.52)	(-6.70)	(-30.67)	(-6.64)	(-1.94)	(-14.36)	(-5.54)
				[-1.8278]	[-0.3132]	[-0.0636]		
Regression	Linear	Linear	Linear	Poisson	Poisson	Poisson	Linear	Linear
Controls	N	Y	Y	N	Y	Y	N	Y
Time FE	Y	Y	Y	Y	Y	Y	Y	Y
Industry FE	N	Y	N	N	Y	N	N	Y
Exchange FE	N	Y	N	N	Y	N	N	Y
Firm FE	N	N	Y	N	N	Y	N	N
Adjusted R^2	4.28%	59.07%	84.90%	5.75%	51.38%	56.99%	58.51%	84.44%
N	848,838	848,838	848,499	848,838	848,799	794,630	678,999	678,360

Panel F: Analyst Coverage

	All Firms						Drop Smallest 20%	
Missing Data	-3.3346 ***	-1.9895 ***	-1.2005 ***	-1.4874 ***	-0.6583 ***	-0.4881 ***	-2.1765 ***	-1.4950 ***
	(-37.38)	(-33.09)	(-22.92)	(-29.55)	(-15.32)	(-14.92)	(-31.47)	(-24.32)
				[-6.7260]	[-2.9770]	[-2.5526]		
Regression	Linear	Linear	Linear	Poisson	Poisson	Poisson	Linear	Linear
Controls	N	Y	Y	N	Y	Y	N	Y
Time FE	Y	Y	Y	Y	Y	Y	Y	Y
Industry FE	N	Y	N	N	Y	N	N	Y
Exchange FE	N	Y	N	N	Y	N	N	Y
Firm FE	N	N	Y	N	N	Y	N	N
Adjusted R^2	5.84%	60.23%	82.98%	6.39%	52.32%	61.01%	59.88%	82.10%
N	848,838	848,838	848,499	848,838	848,777	733,861	678,999	678,360

Table 4: **Individual Characteristic Missingness and Institutional Ownership**

This table reports regressions of institutional ownership on missing data measures for individual characteristics, as in eq. (1). The dependent variable, institutional ownership (FNIO or FNMF), is expressed in percentage points. Missing indicator variables are defined as 1 if the relevant characteristic is missing, and 0 otherwise. T-statistics are reported in parentheses, and standard errors are clustered by firm and date. Controls include: the Koijen and Yogo (2019) market cap instrument, an S&P500 indicator, age, and age squared. Continuous control variables are winsorized at the 1% and 99% levels. Industry is defined as 2-digit SIC code.

Panel A: Investment Characteristics

	All	Insurance		Pension	Mutual Fund	Investment		Mutual
Institution Type:	Institutions	Company	Bank	Fund	Company	Company/Advisor	Miscellaneous	Fund (s12)
Missing Indicator:								
Total Assets	-1.4377 ***	-0.6114 **	-2.0874 ***	-0.7719 **	-1.0994 ***	-0.9402 ***	-0.6510 ***	-0.2331 ***
(AT)	(-7.18)	(-2.34)	(-6.04)	(-2.54)	(-4.87)	(-6.77)	(-3.24)	(-4.08)
Capital	-0.0270	-0.0040	0.2900	0.1711	-0.1078	0.0647	0.0088	-0.0579 **
Expenditures	(-0.28)	(-0.03)	(1.63)	(1.09)	(-0.89)	(0.99)	(0.11)	(-1.98)
(CAPX)								
Total Inventory	-0.0904	-0.2844	-0.2355	-0.0214	-0.1264	-0.0146	0.1487	-0.0576
(INVT)	(-0.48)	(-1.17)	(-0.73)	(-0.08)	(-0.63)	(-0.12)	(0.84)	(-1.05)
Time FE	Y	Y	Y	Y	Y	Y	Y	Y
Industry FE	Y	Y	Y	Y	Y	Y	Y	Y
Exchange FE	Y	Y	Y	Y	Y	Y	Y	Y
Controls	Y	Y	Y	Y	Y	Y	Y	Y
Adjusted R^2	66.09%	71.82%	69.34%	74.63%	66.38%	56.60%	52.76%	59.05%
N	848,838	848,838	848,838	848,838	848,838	848,838	848,838	848,838

Panel B: Valuation Characteristics

	All	Insurance		Pension	Mutual Fund	Investment		Mutual
Institution Type:	Institutions	Company	Bank	Fund	Company	Company/Advisor	Miscellaneous	Fund (s12)
Missing Indicator:								
Stockholder's	-1.6426 ***	-1.6674 ***	-2.5808 ***	-1.6111 ***	-1.6220 ***	-0.9294 ***	-0.4951 ***	-0.3960 ***
Equity (SEQ)	(-17.75)	(-10.50)	(-13.81)	(-8.12)	(-13.49)	(-12.25)	(-7.07)	(-11.34)
Deferred Taxes	0.1516	1.1734 ***	0.7285 ***	1.4069 ***	0.4860 ***	0.0372	-0.0172	0.0956 **
and Investment	(1.29)	(6.31)	(3.07)	(6.35)	(3.34)	(0.37)	(-0.19)	(-1.99)
Tax Credit								
(TXDITC)								
Time FE	Y	Y	Y	Y	Y	Y	Y	Y
Industry FE	Y	Y	Y	Y	Y	Y	Y	Y
Exchange FE	Y	Y	Y	Y	Y	Y	Y	Y
Controls	Y	Y	Y	Y	Y	Y	Y	Y
Adjusted R^2	66.10%	71.86%	69.35%	74.67%	66.39%	56.60%	52.76%	59.06%
N	848,838	848,838	848,838	848,838	848,838	848,838	848,838	848,838

Table 4: **Individual Characteristic Missingness and Institutional Ownership**

Panel C: Profitability Characteristics

Institution Type:	All Institutions	Insurance Company	Bank	Pension Fund	Mutual Fund Company	Investment Company/Advisor	Miscellaneous	Mutual Fund (s12)
Missing Indicator:								
Sales Revenue	-1.4154 ***	-0.1372	-1.7649 ***	0.0609	-1.0162 ***	-0.5980 ***	-0.3215 **	-0.1522 ***
(SALE)	(-9.36)	(-0.62)	(-6.55)	(0.22)	(-5.58)	(-4.77)	(-2.45)	(-2.71)
Interest Expense	-0.0085	-0.4410 ***	-0.2484 *	-0.3288 **	-0.0472	-0.1154 **	-0.0537	-0.0784
(XINT)	(-0.11)	(-3.09)	(-1.83)	(-1.98)	(-0.45)	(-1.98)	(-0.78)	(-2.90)
R&D Expense	-0.4768 ***	-0.5023 ***	-0.7254 ***	-0.5525 ***	-0.5120 ***	-0.3384 ***	-0.3598 ***	-0.1354 ***
(XRD)	(-5.50)	(-4.41)	(-4.61)	(-4.13)	(-5.62)	(-4.81)	(-5.44)	(-4.62)
SG&A Expense	0.0190	-0.3013 **	0.0978	-0.4723 ***	-0.1621	-0.1127	-0.0492	-0.0838 *
(XSGA)	(0.15)	(-1.99)	(0.44)	(-2.76)	(-1.19)	(-1.10)	(-0.49)	(-1.83)
Time FE	Y	Y	Y	Y	Y	Y	Y	Y
Industry FE	Y	Y	Y	Y	Y	Y	Y	Y
Exchange FE	Y	Y	Y	Y	Y	Y	Y	Y
Controls	Y	Y	Y	Y	Y	Y	Y	Y
Adjusted R^2	66.17%	71.85%	69.39%	74.66%	66.43%	56.67%	52.84%	59.11%
N	848,838	848,838	848,838	848,838	848,838	848,838	848,838	848,838

Table 5: **Difference-in-Differences Analysis**

This table reports difference-in-differences regression results, as in eq. (2). The dependent variables are measures of demand, primarily institutional ownership. In Panels A and E, institutional ownership (FNIO) is measured for all institutions in aggregate. In Panels B and C, institutional ownership is measured for individual types and sizes of institutions. 'Inflation-Adj Size Cutoff' and 'EUM-weighted' measures are defined as in Table 3. In Panel D, investor demand is measured via mutual fund ownership (FNMF, s12 file) and analyst coverage. All measures of demand (except analyst coverage) are expressed in percentage points. Treated firms are defined as all financial services firms with no data in the Compustat Point-In-Time database prior to 1993. Compustat classifies firms with SIC codes ranging from 6000 – 6999, excluding codes 6411, 6792, 6794, 6795, as financial services. Untreated firms are defined as all other firms. Post is defined as all periods in and after 1995. Linear regressions are estimated via Ordinary Least Squares. Poisson pseudo-likelihood regressions are estimated as in Correia et al. (2019) and Correia et al. (2020). All regressions are linear unless indicated otherwise. Standard errors are clustered by firm and date, t-statistics are reported in parentheses, and marginal effects for non-linear models are reported in brackets. Continuous control variables are winsorized at the 1% and 99% levels. Controls include: the Koijen and Yogo (2019) market cap instrument, an S&P500 indicator, age, and age squared. Industry is defined as 2-digit SIC code. The sample period covers Q1 1988 – Q4 1999. Both treated and untreated firms are required to be publicly listed in or before Q1 1988. In Panel E, sub-samples are defined based the 'Branching Restrictiveness Index' described in Rice and Strahan (2010), which ranges from values of 0–4, and firms' states are defined based on their address recorded in Compustat. 'High' is an indicator variable defined as one if a firm is located in a state with a Branching Restrictiveness Index value of 3 or 4.

Panel A: Aggregate Institutional Ownership

	All Firms						Drop Smallest 20%	
Treated×Post	1.1561 ***	1.6045 ***	0.9736 ***	0.3391 ***	0.1631 ***	0.1675 ***	1.7180 ***	1.0583 ***
	(4.53)	(9.85)	(6.47)	(8.31)	(7.34)	(6.73)	(9.49)	(6.67)
				[1.6659]	[0.8015]	[0.8334]		
Treated	-1.6701 ***	-1.1162 ***		-0.4301 ***	-0.1350 ***		-1.1873 ***	
	(-6.73)	(-3.99)		(-5.84)	(-3.38)		(-4.00)	
				[-2.1132]	[-0.6631]			
Post	1.1293 ***			0.2122 ***				
	(7.42)			(7.38)				
				[1.0425]				
Regression	Linear	Linear	Linear	Poisson	Poisson	Poisson	Linear	Linear
Controls	N	Y	Y	N	Y	Y	Y	Y
Time FE	N	Y	Y	N	Y	Y	Y	Y
Industry FE	N	Y	N	N	Y	N	Y	N
Exchange FE	N	Y	N	N	Y	N	Y	N
Firm FE	N	N	Y	N	N	Y	N	Y
Adjusted R^2	0.93%	69.89%	95.82%	1.25%	70.49%	75.10%	70.46%	95.75%
N	182,393	182,393	182,215	182,393	182,393	180,095	156,264	156,149

Table 5: **Difference-in-Differences Analysis**

Panel B: Institution Legal Type

Institution Type:	Insurance Company	Bank	Pension Fund	Mutual Fund Company	Investment Company/Advisor	Miscellaneous
Treated×Post	1.7348 ***	1.9132 ***	1.6290 ***	1.2246 ***	0.7050 ***	0.1040
	(6.40)	(6.37)	(5.81)	(6.19)	(5.92)	(0.72)
Controls	Y	Y	Y	Y	Y	Y
Time FE	Y	Y	Y	Y	Y	Y
Firm FE	Y	Y	Y	Y	Y	Y
Adjusted R^2	93.16%	96.12%	93.99%	93.79%	92.93%	86.82%
N	182,215	182,215	182,215	182,215	182,215	182,215

Panel C: Institution Size

EUM Quintile:	Q1 (Small)	Q2	Q3	Q4	Q5 (Large)	30 Largest	Inflation-Adj Size Cutoff	EUM-Weighted
Treated×Post	0.3857 ***	0.4864 ***	0.7232 ***	0.9601 ***	2.4076 ***	5.5155 ***	1.2172 ***	4.1015 ***
	(4.03)	(3.83)	(4.86)	(5.52)	(7.91)	(9.04)	(6.83)	(8.71)
Controls	Y	Y	Y	Y	Y	Y	Y	Y
Time FE	Y	Y	Y	Y	Y	Y	Y	Y
Firm FE	Y	Y	Y	Y	Y	Y	Y	Y
Adjusted R^2	91.33%	92.96%	93.56%	94.79%	95.86%	93.14%	95.94%	94.50%
N	182,215	182,215	182,215	182,215	182,215	182,215	182,215	182,215

Panel D: Mutual Fund Holdings and Analyst Coverage

Dependent Variable:	Mutual Fund Holdings				Analyst Coverage			
Treated×Post	0.2661 ***	0.3715 ***	0.2381 ***	0.2734 ***	1.2529 ***	1.5551 ***	0.8576 ***	0.9666 ***
	(3.74)	(6.54)	(4.83)	(5.16)	(5.55)	(8.52)	(4.96)	(5.27)
Treated	-0.5378 ***	-0.4398 ***			-0.4549 *	-0.2106		
	(-7.77)	(-4.67)			(-1.69)	(-0.67)		
Post	0.1426 ***				0.9062 ***			
	(3.03)				(7.40)			
Sample	All Firms	All Firms	All Firms	Drop Smallest	All Firms	All Firms	All Firms	Drop Smallest
Controls	N	Y	Y	Y	N	Y	Y	Y
Time FE	N	Y	Y	Y	N	Y	Y	Y
Industry FE	N	Y	N	N	N	Y	N	N
Exchange FE	N	Y	N	N	N	Y	N	N
Firm FE	N	N	Y	Y	N	N	Y	Y
Adjusted R^2	0.60%	57.89%	91.77%	91.30%	0.57%	67.88%	92.27%	92.06%
N	182,393	182,393	182,215	156,149	182,393	182,393	182,215	156,149

Panel E: Aggregate Institutional Ownership and State Subgroups

Branching Restrictiveness Index:	0	1	2	3	4	2 ⩽	⩾ 3	All
Treated×Post	1.0955 ***	0.7057 **	1.1781 **	1.3457 ***	0.8788 ***	0.9650 ***	1.0388 ***	0.9088 ***
	(3.90)	(2.50)	(2.30)	(5.42)	(4.02)	(4.63)	(5.78)	(4.72)
High×Treated×Post								0.1443
								(0.69)
Controls	Y	Y	Y	Y	Y	Y	Y	Y
Time FE	Y	Y	Y	Y	Y	Y	Y	Y
Firm FE	Y	Y	Y	Y	Y	Y	Y	Y
Adjusted R^2	96.01%	96.16%	96.86%	95.16%	96.05%	96.34%	95.46%	95.82%
N	30,828	18,982	15,336	52,528	23,622	65,147	117,063	182,215

Table 6: **Institutions' Propensity to Invest in Firms with Missing Data**

This table reports regressions of the fraction of institutions' portfolios invested in firms with no Compustat data on a variety of institution characteristics, as in eq. (3) and (4). Dependent variables are defined as the 'Fraction of EUM,' equal to the fraction of an institution's equity under management invested in firms with no Compustat coverage, or the 'Fraction of Firms,' defined as the fraction of firms that the institution is invested in with no Compustat coverage. Panel A focuses on 13f institutions, and observations are recorded at the institution-quarter frequency. The sample in Panel A is 1986–2021. Panel B focuses on the individual mutual funds from the s12 data, and observations are recorded at the mutual fund-quarter frequency. The sample in Panel B is 1980–2009.

Panel A: Institutional Investors

Dependent Variable:	Fraction of EUM			Fraction of Firms		
Log(EUM)	-0.0181	-0.0201	0.0100	0.0097	0.0058	0.0303
	(-1.24)	(-1.38)	(0.40)	(0.74)	(0.44)	(1.31)
Age	0.0012 **	0.0013 **	-0.0141 *	-0.0003	-0.0001	-0.0176 ***
	(2.06)	(2.17)	(-1.70)	(-0.70)	(-0.21)	(-3.30)
Portfolio Turnover	0.5176 ***		0.1708 **	0.6827 ***		-0.0086
	(3.66)		(2.07)	(3.72)		(-0.13)
Turnover: Quint 1		-0.2742 ***			-0.4218 ***	
		(-3.20)			(-3.74)	
Turnover: Quint 2		-0.2706 ***			-0.3988 ***	
		(-3.81)			(-3.50)	
Turnover: Quint 3		-0.2572 ***			-0.3412 ***	
		(-3.52)			(-3.06)	
Turnover: Quint 4		-0.1482 ***			-0.1344 *	
		(-2.43)			(-1.75)	
Insurance	-0.2679 ***	-0.2590 ***		-0.2413 **	-0.2178 **	
	(-3.04)	(-2.95)		(-2.41)	(-2.22)	
Bank	-0.1922 **	-0.1742 *		-0.3854 ***	-0.3321 ***	
	(-2.06)	(-1.86)		(-4.60)	(-4.22)	
Pension Fund	-0.3154 ***	-0.3012 ***		-0.4354 ***	-0.3945 ***	
	(-3.49)	(-3.35)		(-4.50)	(-4.24)	
Mutual Fund Co.	-0.0243	-0.0199		0.0932	0.0941	
	(-0.28)	(-0.23)		(0.98)	(1.01)	
Investment Co.	0.0933	0.0944		0.0796	0.0831	
	(1.21)	(1.22)		(1.10)	(1.17)	
Time FE	Y	Y	Y	Y	Y	Y
Institution FE	N	N	Y	N	N	Y
Adjusted R^2	3.72%	3.74%	30.52%	8.50%	8.64%	35.58%
N	324,900	324,900	324,474	324,900	324,900	324,474

Table 6: **Institutions' Propensity to Invest in Firms with Missing Data**

Panel B: Mutual Funds

Dependent Variable:	Fraction of EUM			Fraction of Firms		
Log(EUM)	-0.0079	0.0062	-0.0961 ***	0.0261 *	0.0409 ***	-0.0784 **
	(-0.73)	(0.63)	(-3.36)	(1.97)	(3.06)	(-2.44)
Age	-0.0011	-0.0008	-0.0015	-0.0015 *	-0.0012	0.0014
	(-1.50)	(-1.14)	(-0.45)	(-1.80)	(-1.46)	(0.33)
Portfolio Turnover	1.6204 ***		0.8475 ***	1.8806 ***		0.7896 ***
	(5.04)		(4.66)	(4.86)		(3.91)
Turnover: Quint 1		-0.5294 ***			-0.6198 ***	
		(-5.49)			(-5.30)	
Turnover: Quint 2		-0.3743 ***			-0.4305 ***	
		(-4.44)			(-4.25)	
Turnover: Quint 3		-0.3008 ***			-0.3372 ***	
		(-4.15)			(-3.91)	
Turnover: Quint 4		-0.2173 ***			-0.2336 ***	
		(-3.60)			(-3.42)	
Active Share	1.5848 ***		0.2637	1.8806 ***		0.4340
	(7.06)		(1.23)	(6.80)		(1.64)
Active Share: Quint 2		0.1860 ***			0.2299 ***	
		(4.93)			(4.63)	
Active Share: Quint 3		0.3914 ***			0.4790 ***	
		(6.27)			(5.92)	
Active Share: Quint 4		0.6955 ***			0.8383 ***	
		(7.75)			(7.26)	
Active Share: Quint 5		1.0967 ***			1.2343 ***	
		(7.87)			(7.56)	
Index Fund	1.0719 ***	0.5149 ***		1.2721 ***	0.6236 ***	
	(5.72)	(4.62)		(5.76)	(4.64)	
Enhanced Index	0.7968 ***	0.4503 ***		0.9607 ***	0.5535 ***	
	(5.58)	(5.73)		(5.67)	(5.74)	
Time FE	Y	Y	Y	Y	Y	Y
Mutual Fund FE	N	N	Y	N	N	Y
Adjusted R^2	25.29%	27.59%	43.55%	31.26%	33.24%	48.78%
N	79,712	79,712	79,625	79,712	79,712	79,625

Table 7: **Earnings Announcements and Missing Data**

This table reports regressions of abnormal returns measured during and after quarterly earnings announcements on an indicator for missing Compustat data. Regressions are estimated as in eq. (5), standard errors are clustered by time, and t-statistics are reported in parentheses. 'Missing Data' is an indicator variable defined as 1 if a firm does not have any Compustat data available in the most recent fiscal year-end. The dependent variables are defined as absolute cumulative abnormal returns ('ACAR') around quarterly earnings announcements. Trading day $\tau = 0$ is defined as the earnings announcement date. In Panel A, ACARs are constructed over the window $\tau = [-1, 1]$. In Panel B, post-announcement ACARs are constructed over the window $\tau = [2, 60]$. All observations are recorded at the firm-announcement frequency, and controls are measured as of the most recent quarter-end. Controls in all regressions include: analyst coverage, institutional ownership (FNIO), stock beta, log market cap, an S&P500 indicator, prior 1-year return, prior return measured over years -5:-1, net stock issuance, share turnover, and age. Continuous control variables are winsorized at the 1% and 99% levels. Industry is defined as 2-digit SIC code. The sample period covers 1980–2021.

Panel A: Earnings Announcement Returns, ACAR[-1,1]

Missing Data	0.5372 ***	0.3566 ***	0.4426 ***	0.3157 ***	0.9743 ***	1.1653 ***	0.7893 ***	1.0097 ***
	(5.38)	(3.37)	(5.03)	(3.22)	(8.06)	(8.97)	(7.42)	(8.41)
Missing Data×	-0.0745 ***	-0.0616 ***			-0.1301 ***	-0.1553 ***		
Analyst Coverage	(-4.48)	(-3.50)			(-6.50)	(-6.29)		
Missing Data×			-0.0396 ***	-0.0569 ***			-0.0564 ***	-0.1305 ***
FNIO			(-3.98)	(-4.44)			(-4.39)	(-4.85)
Normal Return	Market	Market	Market	Market	FF3	FF3	FF3	FF3
Controls	Y	Y	Y	Y	Y	Y	Y	Y
Time FE	Y	Y	Y	Y	Y	Y	Y	Y
Industry FE	Y	N	Y	N	Y	N	Y	N
Exchange FE	Y	N	Y	N	Y	N	Y	N
Firm FE	N	Y	N	Y	N	Y	N	Y
Adjusted R^2	10.31%	14.92%	10.31%	14.92%	10.72%	15.71%	10.72%	15.71%
N	586,970	586,319	586,970	586,619	550,779	550,321	550,779	550,321

Panel B: Post-Announcement Drift, ACAR[2,60]

Missing Data	1.0242 ***	0.8523 ***	0.8502 ***	0.6862 ***	0.6677 **	1.0716 **	0.3788	0.8167 **
	(3.89)	(2.81)	(3.70)	(2.57)	(2.01)	(2.37)	(1.32)	(2.06)
Missing Data×	-0.0635	-0.0887 *			-0.1007 *	-0.1656 **		
Analyst Coverage	(-1.44)	(-1.79)			(-1.87)	(-2.26)		
Missing Data×			0.0212	-0.0255			0.0530	-0.0836
FNIO			(0.73)	(-0.84)			(1.39)	(-1.39)
Normal Return	Market	Market	Market	Market	FF3	FF3	FF3	FF3
Controls	Y	Y	Y	Y	Y	Y	Y	Y
Time FE	Y	Y	Y	Y	Y	Y	Y	Y
Industry FE	Y	N	Y	N	Y	N	Y	N
Exchange FE	Y	N	Y	N	Y	N	Y	N
Firm FE	N	Y	N	Y	N	Y	N	Y
Adjusted R^2	13.07%	18.27%	13.07%	18.27%	13.46%	18.66%	13.46%	18.66%
N	586,970	589,619	586,970	589,619	550,779	550,321	550,779	550,321

Table 8: **Information Assimilation and Missing Data**

This table reports regressions of various measures of informational efficiency on an indicator for missing Compustat data, as in eq. (5). 'Missing Data' is an indicator variable defined as 1 if a firm does not have any Compustat data available in the most recent fiscal year-end. In Panel A, the dependent variables are defined as measures of return autocorrelations, estimated as the absolute coefficient value ($|\rho|$), t-statistic ($|\frac{\rho}{se(\rho)}|$, or r-squared from the regression in eq. (6). In Panel B, the dependent variables are defined as measures of price delay, estimated as in eq. (8), (9), or (10). All observations are recorded at the annual frequency. Autocorrelation and price delay measures are constructed using return data from July in year t through June in year $t+1$. Missing Compustat data is measured as of the $t-1$ fiscal year-end. Controls are measured as of the end of June in year t. Standard errors in all regressions are clustered by time, and t-statistics are reported in parentheses. Controls in all regressions include: analyst coverage, institutional ownership (FNIO), stock beta, log market cap, an S&P500 indicator, prior 1-year return, prior return measured over years -5:-1, net stock issuance, share turnover, and age. Continuous control variables are winsorized at the 1% and 99% levels. Industry is defined as 2-digit SIC code. The sample period is 1980–2021.

Panel A: Daily Return Autocorrelation

Dependent Variable:	$\|\rho\|$	$\|\rho\|$	$\|\frac{\rho}{se(\rho)}\|$	$\|\frac{\rho}{se(\rho)}\|$	$R^2(\%)$	$R^2(\%)$
Missing Data	0.0208 **	0.0218 **	0.3232 **	0.3420 **	1.4677 ***	1.5339 ***
	(2.25)	(2.27)	(2.02)	(2.06)	(2.79)	(2.80)
Missing Data×	0.0002		0.0098		-0.0466	
Analyst Coverage	(0.22)		(0.59)		(-0.82)	
Missing Data×		-0.0009		-0.0120		-0.1177
FNIO		(-0.67)		(-0.57)		(-1.41)
Controls	Y	Y	Y	Y	Y	Y
Time FE	Y	Y	Y	Y	Y	Y
Industry FE	Y	Y	Y	Y	Y	Y
Exchange FE	Y	Y	Y	Y	Y	Y
Adjusted R^2	29.42%	29.42%	26.16%	26.16%	28.62%	28.64%
N	190,559	190,559	190,559	190,559	190,559	190,559

Panel B: Price Delay

Dependent Variable:	D1 (%)	D1 (%)	D2	D2	D3	D3
Missing Data	4.6485 ***	4.6829 ***	0.2066 ***	0.2026 ***	0.1559 ***	0.1591 ***
	(4.97)	(5.29)	(3.84)	(3.87)	(2.98)	(3.19)
Missing Data×	-0.4511 **		-0.0249 **		-0.0278 **	
Analyst Coverage	(-2.60)		(-2.35)		(-2.43)	
Missing Data×		-0.4521 *		-0.0184 **		-0.0291 **
FNIO		(-1.91)		(-2.28)		(-2.61)
Controls	Y	Y	Y	Y	Y	Y
Time FE	Y	Y	Y	Y	Y	Y
Industry FE	Y	Y	Y	Y	Y	Y
Exchange FE	Y	Y	Y	Y	Y	Y
Adjusted R^2	36.51%	36.51%	4.32%	4.32%	3.94%	3.94%
N	194,365	194,365	194,365	194,365	194,365	194,365

Appendix Table 1: **Characteristic Definitions**

Panel A: Compustat Characteristics

Variable	Acronym	Freq	Description
Asset Growth	a_growth	A	The percentage change in total assets (AT) over the preceding year.
	a_growthq	Q	The percentage change in total assets (ATQ) over the preceding quarter.
Accruals	acc	A	The change in current assets minus the change in cash and short-term investments minus the change in current liabilities plus the change in debt in current liabilities plus the change in income taxes payable minus the change in depreciation and amortization, scaled by the average of current and lagged total assets $(\Delta ACT - \Delta CHE - \Delta LCT + \Delta DLC + \Delta TXP - \Delta DP) \, / \left(\frac{AT_{t-1}+AT_t}{2}\right)$.
	accq	Q	$(\Delta ACTQ - \Delta CHEQ - \Delta LCTQ + \Delta DLCQ + \Delta TXPQ - \Delta DPQ) \, / \left(\frac{ATQ_{t-1}+ATQ_t}{2}\right)$
Book Equity	be	A	The book value of equity (BE), defined as shareholders' equity (SH) plus deferred taxes (txditc) minus preferred stock (PS). SH is equal to shareholders' equity (SEQ). If missing, SH is equal to the sum of common equity (CEQ) and preferred stock (PS). If also missing, SH is the equal to the difference between total assets (AT) and total liabilities (LT). Depending on availability, PS is redemption value (item PSTKRV), liquidating value (item PSTKL), or par value (item PSTK). BE can take negative values.
	beq	Q	The book value of equity (BEQ), defined as shareholders' equity (SHQ) plus deferred taxes (TXDITCQ) minus preferred stock (PSQ). SHQ is equal to shareholders' equity (SEQQ). If missing, SHQ is equal to the sum of common equity (CEQQ) and preferred stock (PSQ). If also missing, SHQ is the equal to the difference between total assets (ATQ) and total liabilities (LTQ). PSQ is equal to PSTKQ when available. BEQ can take negative values.
Book-to-Market	beme	A	The book value of equity (BE) scaled by the market value of equity. Market equity is measured as of the prior December-end.
	bemeq	Q	The book value of equity (BEQ) scaled by the market value of equity. Market equity is measured as of the prior quarter-end.
Growth in CAPX	capx_growth	A	The ratio of the change in capital expenditures (CAPX).
Cash Profit	cp	A	Total sales revenue minus cost of goods sold, SG&A expense, and interest expense plus R&D expense minus the Annual change in accounts receivable minus the annual change in inventory minus the annual change in prepaid expenses plus the annual change in deferred revenue plus the annual change in trade accounts receivable plus the annual change in accrued expenses, scaled by book equity $[SALE - COGS - XSGA - XINT + XRD - \Delta RECT - \Delta INVT - \Delta XPP + \Delta(DRC + DRLT) + \Delta AP + \Delta XACC]/BE$.
Net Debt-to-Price	debt_price	A	Net debt (DLTT + DLC + PSTK + DVPA - TSTKP - CHE), scaled by the market value of equity measured as of the prior December-end.
Dividend Yield	div_yield	A	Total dividends (DVT) scaled by the market value of equity. Market equity is measured as of the prior December-end.
Earnings-to-Price	earnings_price	A	Earnings (IB) scaled by the market value of equity. Market equity is measured as of the prior December-end.
	earnings_priceq	Q	IBQ scaled by the market value of equity measured as of the prior quarter-end.

Appendix Table 1: **Characteristic Definitions**

Variable	Acronym	Freq	Description
F-score	f_score	A	F-score $= 1_{IB>0} + 1_{\Delta ROA>0} + 1_{CFO>0} + 1_{CFO>IB} + 1_{\Delta DTA<0\|DLTT=0\|DLTT_{-12}=0} + 1_{\Delta ATL>0} + 1_{EqIss\leqslant 0} + 1_{\Delta GM>0} + 1_{\Delta ATO>0}$, where IB is income before extraordinary items, ROA is income before extraordinary items scaled by lagged total assets, CFO is cash flow from operations, DTA is total long-term debt scaled by total assets, DLTT is total long-term debt, ATL is total current assets scaled by total current liabilities, EqIss is the annual difference between total shares outstanding ('shares'), GM equals one minus the ratio of cost of goods sold and total sales revenues, and ATO equals total sales revenue, scaled by total assets.
Gross Profit	gp	A	Total sales revenue (SALE) minus cost of goods sold (COGS), scaled by total assets (AT).
	gpq	Q	(SALEQ - COGSQ) / ATQ.
Investment-to-Capital	inv_cap	A	Capital expenditures (CAPX) divided by property, plant, and equipment (PPENT).
Investment Growth	inv_growth	A	The change in property, plant, and equipment plus the change in inventory, scaled by lagged total assets (ΔPPENT + ΔINVT)/(AT_{t-1}).
	inv_growthq	Q	(ΔPPENTQ + ΔINVTQ)/(ATQ_{t-1}).
Market Leverage	leverage	A	Total assets scaled by the market value of equity, AT / ME.
	leverageq	Q	ATQ / QT_ME.
Net Operating Assets	noa	A	Total debt (DLC + DLTT) plus minority interest plus total preferred stock plus the book value of common equity minus cash and short-term investments, scaled by lagged total assets (DLC + DLTT + MIB + PS + CEQ - CHE)/(AT_{t-1}). PS is defined as in the Book Equity definition above.
	noaq	Q	(DLCQ + DLTTQ + MIBQ + PSQ + CEQQ - CHEQ)/(ATQ_{t-1}). PSQ is defined as in the Book Equity definition above.
O-Score	o_score	A	O-score $= -(-1.32 - 0.407 * log(ADJASSET/CPI) + 6.03 * TLTA - 1.43 * WCTA + 0.076 * CLCA - 1.72 * OENEG - 2.37 * NITA - 1.83 * FUTL + 0.285 * INTWO - 0.521 * CHIN)$, where ADJASSET is adjusted total assets equal to total assets plus 10% of the difference between book equity and market equity $AT + 0.1 * (ME - BE)$. CPI is the consumer price index. TLTA is equal to book value of debt $(DLC + DLTT)$ divided by ADJASSET. WCTA is current assets minus current liabilities scaled by adjusted assets $(ACT - LCT)/ADJASSET$. CLCA is current liabilities divided by current assets LCT/ACT. OENEG is a dummy equal to 1 if total liabilities exceed total assets $1(LT > AT)$. NITA is net income over assets IB/AT. FUTL is pre-tax income over total liabilities PT/LT. INTWO is a dummy equal to one if net income is negative for the current and prior fiscal year $1(MAX[IB_t, IB_{t-1}] < 0)$. CHIN is the change in net income defined as $(IB_t - IB_{t-1})/(\|IB_t\| + \|IB_{t-1}\|)$.
	o_scoreq	Q	Defined as above, using quarterly values of total assets (ATQ), book equity (BEQ),total debt (DLCQ and DLTTQ), current assets (ACTQ), current liabilities (LCTQ), total liabilities (LTQ), and net income (IBQ).
Operating Profit	op	A	Total sales revenue (SALE) minus cost of goods sold (COGS), SG&A expense (XSGA), and interest expense (XINT), scaled by book equity.
	opq	Q	(SALEQ - COGSQ - XSGAQ - XINTQ) / BEQ.

Appendix Table 1: **Characteristic Definitions**

Variable	Acronym	Freq	Description
Operating Profit Adjusted for R&D Expense	op_rd	A	Total sales revenue (SALE) minus cost of goods sold (COGS), SG&A expense (XSGA), and interest expense (XINT) plus R&D expense (XRD), scaled by book equity.
	op_rdq	Q	(SALEQ - COGSQ - XSGAQ - XINTQ + XRDQ) / BEQ.
Return on Assets	roa	A	Net income (IB) scaled by total assets (AT).
	roaq	Q	IBQ / ATQ.
Return on Equity	roe	A	Net income (IB) scaled by book equity (BE).
	roeq	Q	IBQ / BEQ.
Sales-to-Price	sales_price	A	Sales (SALE) scaled by the market value of equity. Market equity is measured as of the prior December-end.
	sales_priceq	Q	SALEQ / QT_ME.
Sales Growth	sg	A	The annual percentage change in SALE.
	sgq	Q	Quarterly percentage change in SALE.
Shares	shares	A	Total shares outstanding, measured as of the end of the prior December.
	sharesq	Q	Total shares outstanding, measured as of the end of the prior quarter.
Tobin's Q	tobin_q	A	Total assets (AT) plus the market value of equity from the prior Demeber end (ME) minus cash and short-term investments (CEQ) minus deferred taxes (TXDB), scaled by total assets (AT).
	tobin_qq	Q	(ATQ + MEQ - CEQQ - TXDBQ) / ATQ
Z-score	z_score	A	Z-score $= 1.2 * (ACT - LCT)/AT + 1.4 * (RE/AT) + 3.3 * (NI + XINT + TXT)/AT + 0.6 * (ME/LT) + SALE/AT$, where ACT is current assets, LCT is current liabilities, AT is total assets, RE is retained earings, NI is net income, XINT is interest expense, TXT is total taxes, ME is the market value of equity, LT is total liabilities, and SALE is total sales revenue.
	z_scoreq	Q	$1.2 * (ACTQ - LCTQ)/ATQ + 1.4 * (REQ/ATQ) + 3.3 * (NIQ + XINTQ + TXTQ)/ATQ + 0.6 * (QT\_ME/LTQ) + SALEQ/ATQ$.
Panel B: Non-Compustat Firm Characteristics			
Analyst Coverage	analyst_coverage	Q	The number of analysts covering a firm in a given quarter, equal to the number of quarterly earnings forecasts made by unique analysts ('NUMEST' from the I/B/E/S Summary file).
Beta	beta	M	Monthly CAPM beta estimated using the prior 60 months of returns; require a minimum of 36 months of data.
Fractional Number of Institutional Owners	FNIO	Q	The total number of institutions that own shares of a firms in a given quarter, scaled by the total number of institutions in the 13f data in that quarter.
Fractional Number of Mutual Fund Owners	FNMF	Q	The total number of mutual funds that own shares of a firms in a given quarter, scaled by the total number of mutual funds in the 13f data (s12 file) in that quarter.

Appendix Table 1: **Characteristic Definitions**

Variable	Acronym	Freq	Description
Fraction of Shares Held by Institutions	FSIO	Q	The fraction of a firms' total shares outstanding held by institutions in a given quarter.
Fraction of Shares Held by Mutual Funds	FSMF	Q	The fraction of a firms' total shares outstanding held by mutual funds in a given quarter.
Market Cap Instrument	KY_me	Q	Constructed as in Koijen and Yogo (2019), and equal to the log of the counter-factual market equity if all institutions held an equally-weighted portfolio of their investable universe. Each institution's investable universe is defined as all stocks that they currently hold or have held over the prior three years. Institutions' counter-factual investments are defined as total equity under management multiplied by 1/N, where N is the total number of firms in the institutions' investable universe. Each firm's counter-factual market equity is defined as the sum of all counter-factual investments for each institution.
Net Issuance	net_iss	A	Annual log change in split-adjusted shares outstanding. Shares outstanding are measured as of the prior December-end.
Number of Institutional Owners	NIO	Q	The total number of institutions that own shares of a firms in a given quarter.
Turnover	turnover	A, Q	Trading volume, averaged over the prior M months, divided by total shares outstanding. In annual regressions, $M = 12$. In quarterly regressions, $M = 3$.
			Panel C: Institution and Mutual Fund Characteristics
Active Share	active share	Q	Constructed as in Cremers and Petajisto (2009) and Petajisto (2013), and equal to the sum of absolute differences between each mutual fund's portfolio weights and the fund's benchmark's portfolio weights. Active share data is obtained from Annti Petajisto's website, and is available only for the mutual fund (s12) data.
Age	age	Q	The number of quarters that the institution (s34 data) or mutual fund (s12 data) has appeared in the 13f database.
Portfolio Turnover	portfolio turnover	Q	Constructed following Yan and Zhang (2009), and equal to the four-quarter average of an institution's or mutual fund's churn rate. Churn rate is equal to the minimum of aggregate purchases and aggregate sales, both measured over the most recent quarter, scaled by the average of equity under management at the beginning and end of the quarter.