THE UNIVERSITY OF CHICAGO

THE EFFECTS OF ALTERNATIVE DATA ON MISREPORTING

A DISSERTATION SUBMITTED TO

THE FACULTY OF THE UNIVERSITY OF CHICAGO

BOOTH SCHOOL OF BUSINESS

IN CANDIDACY FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

BY

JONAS PIERO DALMAZZO

CHICAGO, ILLINOIS

AUGUST 2025

Abstract

New data sources, such as transactional data from point-of-sale systems and store traffic data from satellite imagery, help reduce information asymmetry between managers and external stakeholders. I investigate whether these new sources of information influence managers' decisions to manipulate real activities and accounting reports, and whether this is driven by changes in investors' monitoring capabilities. To explore the impact of the adoption and use of alternative data by financial markets, I develop a novel measure based on employment data that captures sophisticated investors' investment in alternative data. My findings indicate that increased coverage of alternative data affects managerial behavior in treated firms, reducing the likelihood of real earnings management and leading to a shift toward some more aggressive accounting practices.

Dedication & Acknowledgments

To my wife Madeleine, for your endless support, patience, and belief in me throughout this journey.

I thank my committee members—Philip Berger, Christian Leuz, Delphine Samuels, and Joao Granja—for their mentorship and support. I thank Fabian Nagel, Maria Khrakovsky, Pietro Ramella, Roope Keloharju, Matthias Breuer, Anna Costello, Charles McClure, Maximilian Muhn, Thomas Rauter, Christopher Stewart, Nancy Lixin Su (discussant), Anastasia Zakolyukina, and all the seminar participants at the University of Chicago, INSEAD, HEC Paris, Boston College, Bocconi University, SMU Cox School of Business, and the Emerging Scholar Conference - Frankfurt School of Management for their helpful suggestions and feedback.

This research was funded in part by the John and Serena Liew Fellowship Fund at the Fama-Miller Center for Research in Finance, University of Chicago Booth School of Business. I gratefully acknowledge the financial support of the Stevens Doctoral Program and the Chookaszian Accounting Research Center.

Researcher(s)' own analyses calculated (or derived) based in part on (i) retail measurement/consumer data from Nielsen Consumer LLC ("NielsenIQ"); (ii) media data from The Nielsen Company (US), LLC ("Nielsen"); and (iii) marketing databases provided through the respective NielsenIQ and the Nielsen Datasets at the Kilts Center for Marketing Data Center at The University of Chicago Booth School of Business.

The conclusions drawn from the NielsenIQ and Nielsen data are those of the researcher(s) and do not reflect the views of Nielsen. Nielsen is not responsible for, had no role in, and was not involved in analyzing and preparing the results reported herein.

Table of Contents

List of Figures

List of Tables

1 Introduction

In recent years, the amount of data available to capital market participants has grown significantly, with new sources, such as satellite measurement of store traffic and transactional data collected at the point of sale, providing more precise and frequent information about firms' performance. These alternative data sources act as market-based proxies for firm revenues, allowing for timely assessment of key profitability drivers such as consumer demand and sales. By reducing the information asymmetry between managers and external stakeholders, alternative data have reshaped public firms' information environments. This shift has improved the accuracy of short-term forecasts, at the expense of long-term ones (Dessaint, Foucault, and Fresard (2024)), and reduced opportunistic trading by managers (Zhu (2019)). To explore the implications of these changes, I focus on alternative data used to measure real-time revenues, including satellite coverage of retailers' parking lots and point-of-sale transaction data. I investigate whether the introduction of these new signals enhances monitoring by investors and whether this affects managers' misreporting and manipulation of real activities.

The short-termism of managers at public firms has often been raised as a concern. And according to Graham, Harvey, and Rajgopal (2005), almost all managers of US public companies do face pressure to meet or beat short-term targets, and many strive to do this, irrespective of the long-term value consequences. Based on experimental evidence, Cade, Gunn, and Vandeberg (2024) report that nearly 30% of executives of U.S. public companies have changed operational activities to meet or beat their targets, despite future performance concerns. Empirically, the literature on short-termism, including Terry (2023), suggests managers may act inefficiently to avoid incurring small losses, which then harms their firms' long-term performance.

Whether a frequently observed signal of firm performance will reduce this short-term bias is not obvious. Given the evidence of frequent disclosure requirements aggravating managerial short-termism through a higher prevalence of real activity manipulation (REM) (Gigler, Kanodia, Sapra, and Venugopalan (2014); Ernstberger, Link, Stich, and Vogler (2017); Kraft, Vashishtha, and Venkatachalam (2018)), the question of how new data sources affect managerial actions and misreporting is both pertinent and timely. The alternative data examined here differ from firm financial reports in that they do not originate from within the firm. Quarterly and annual results are affected by managerial incentives through managers' discretion in compiling the reports, but the data from outside the firm provide a potentially less biased signal. The continuous observation of a firm's transactions (through data provided by third-party vendors), for example, enables sophisticated investors to form more precise expectations of several key indicators, such as profitability and revenues.

I thus investigate whether alternative data enhance the ability of external stakeholders to monitor firms' performance and consequently discipline managers. Monitoring by better-informed stakeholders can diminish the effectiveness of short-term strategies aimed at meeting or beating performance targets (i.e., short-termism). Specifically, the type of short-term actions examined here, following Roychowdhury (2006), are overproduction to reduce the cost of goods sold, aggressive discounting to temporarily boost revenues, and cuts to discretionary spending, such as research and development (R&D) and advertising, to reduce expenses.

A critical distinction between real and accrual earnings manipulations lies in the timing and order in which these actions are potentially taken (Zang (2012); Cohen and Zarowin (2010)). Myopic actions, due to the nature of the operational changes involved, occur throughout the year and can require more time for implementation. In contrast, accrual earnings manipulation (AEM), such as drawing down reserves, can typically be realized quickly. I focus on real activity manipulation because of the timing of the action as well as its connection to the information that alternative data provide: revenue boosts from aggressive discounting may be easier to detect through transaction data than understated bad debt expenses. Whether a manager will resort to AEM to counteract the decreased ability to pursue REM will depend on the perceived costs and benefits of the action as well as the strength of the incentives.[1]

I use two distinct settings to study the effects of alternative data on managerial behavior: transaction data and satellite coverage of retailers. The first setting is based on transaction data collected at the point-of-sale, such as retail scanner data. As the market for alternative data has ballooned, more sources of data have emerged. To be able to study their effects, I introduce a novel measure of adoption of alternative data by sophisticated investors based on human capital acquisition. This new labor-based measure is built by assessing the skills of employees hired by sophisticated investors.

[1]The introduction of data that enhance external monitoring might not directly affect the compensation and incentives of the managers. If, following the increased monitoring, managers pursue less REM (such as cutting R&D), they may instead opt for more aggressive accounting if the perceived benefit of beating the target exceeds the costs of misreporting. According to Cade et al. (2024), REM appears significantly more prevalent than AEM, and thus the extent to which firms will re-allocate misreporting to aggressive accounting from real activity manipulation is unclear a priori.

An increase in adoption of new data by the financial industry, proxied by the hiring of workers with skills related to Big Data, should be reflected in a higher investment in these sources of information, leading to higher utilization of these signals. Relying on adoption, instead of availability, helps mitigate the issue of tracking the various sources of alternative data. As noted by Dessaint et al. (2024), between 2008 and 2017, the number of alternative data providers increased substantially. Measuring the adoption of alternative data based on investment in human capital should proxy for the level of interest in these signals. To identify which firms are covered by alternative data, I rely on a hand-collected list of firms present in a source of transaction-level data.

The second setting investigates the effects of satellite data introduced for commercial and trading purposes. These satellite data measure the volume of clients visiting a retailer's store, in a timely fashion, by showing the use of parking lots (see Zhu (2019); Katona, Painter, Patatoukas, and Zeng (2024); Bonelli and Foucault (2023)). I rely on data provided by RS Metrics, which contain information on which retailers were covered (and since when) in satellite data sold to investors. The use of this second setting allows me to provide additional evidence on the effects of alternative data on misreporting and REM. I choose this setting for two reasons. First, it enables me to test the core research question in a different empirical context. Second, by restricting the analysis to a specific sector (retail), I can exploit a more clearly timed introduction of alternative data. This sharper rollout allows me to deploy a more canonical difference-in-differences testing framework, which improves identification. Though useful, this secondary setting does have some limitations. First, the sample size is smaller, and, second, the REM metrics derived from the literature may be less informative in the retail context.

Before examining whether the introduction of the new data influences managerial behavior, I test the usefulness of these signals in inferring potential real activity manipulation. If the new information available to sophisticated investors contributes to their monitoring of short-termism, it should be possible to measure a change in capital market behavior for firms exhibiting signs of manipulation. To test this hypothesis, I focus on shorting in the weeks between the fiscal year-end date and the announcement of the results. During this period, the firm's economic activities have been realized, but the market remains uninformed about its accounting results, which allows sophisticated investors to gain a meaningful informational advantage by observing the proxies studied here.

Short positions reflect the behavior of sophisticated investors (given the nature of the data covering institutional funds), and prior evidence indicates that shorting discourages AEM (Massa, Zhang, and Zhang (2015)) and REM (Cai, Guo, and Tan (2024)). Moreover, shorting is costly and is typically undertaken only when an investor strongly expects a stock's future downturn. Upon detecting signs of managerial myopia, investors may choose to short the stock for one of two reasons: they either view the myopia as a signal of deteriorating performance or believe that it could lead to subsequent declines in performance (Kothari, Mizik, and Roychowdhury (2016); Cohen and Zarowin (2010)). Using announced results to construct a measure of manipulation,[2] I examine whether sophisticated investors respond to potential earnings manipulation (realized in the prior fiscal year) during the period

[2]The REM proxies are over-production, cuts to discretionary expenditures, and aggressive discounting reflected in abnormally low cash flows. These measures are based on Roychowdhury (2006) and have been used in the literature to detect inefficient managerial actions, as in Irani and Oesch (2016). Similarly, a measure of discretionary accruals is built based on the modeling of Dechow, Sloan, and Sweeney (1995).

between the fiscal year-end and the subsequent earnings announcement.

In the transaction data setting, I find that firms covered by the new data sources face stronger shorting during the weeks between the end of fiscal year and the earnings announcement when they show signs of real activity manipulation, such as aggressive discounting, but not when they show abnormal discretionary accruals. These results are robust to controlling for realized performance, such as ROA, net income, and earnings surprise. I interpret them as an increase in monitoring by sophisticated investors. Additionally, treated firms showing stronger signs of real activity manipulation have more negative post-announcement returns, which provides a monetary incentives for investors to monitor. The focus on the period between the end of the year and the announcement of results indicates that the alternative data studied can identify short-termism before the publication of the financial statements. I repeat the analysis in the satellite data setting for robustness. The results, although inconsistently significant, suggest that the monitoring mechanism is present as well. Together, the analyses show that financial markets can better monitor inefficient managerial actions when newly available data illuminate firms' real performance.

Ideally, the test of the monitoring mechanism at the core of this paper would rely on observing the content of different sources of alternative data, like what sophisticated investors observe. However, that is infeasible. To mitigate concerns related to the use of the real activity manipulation measures computed from the announced results, I use detailed data on advertising expenditures at the ad-date level. The use of these granular data allow me to delve into whether changes in advertising induce a reaction in financial markets, and I use this as a proxy for monitoring. If sophisticated investors do use these (or similar) data to infer firm performance, it would be reasonable to expect a capital market reaction to changes in expenditure patterns that may signal manipulation or poor performance, as discussed above.

I show that drops in advertising expenditures lead to a negative market reaction, with a standard deviation drop in monthly advertising expenditures raising the short interest index (SII) approximately 1.7%. These findings are accompanied by a decrease in trading volume and an increase in bid-ask spread. The results suggest that managers might not cut expenditures to the extent described in survey evidence (Graham et al. (2005)) and previous papers (Cohen, Mashruwala, and Zach (2010)), as negative changes in advertising patterns appear to lead to adverse reactions by sophisticated investors. Given the findings of Shapiro, Hitsch, and Tuchman (2021) on the average negative return on investment in advertising, it is difficult to assess whether a drop in expenditures can be a correction to over-investing or a sign of the firm cutting an expense as it tries to meet an earnings target. Even so, the results can be interpreted as sophisticated investors using advertising expenditure data like they use other novel sources of data to monitor firms. The validity of the monitoring mechanism enabled by advertising data complements the findings on satellite coverage and transaction data.

Next, I study whether the stronger monitoring is associated with changes in firm operations. I estimate whether REM indicators and accruals manipulation show evidence of a change in behavior. I first use the adoption of transaction data to study whether the increased monitoring is reflected in a change of managers' behavior. I provide evidence of some decrease in overproduction and an increase in discretionary expenses (such as R&D),

together with an increase in cash flows from operations. The results indicate an overall decrease of myopic actions. The decrease in real activity manipulation is accompanied by an increase in total accruals, which could indicate some substitution toward more aggressive accounting. I also study the consequences of accounting misreporting via Accounting and Auditing Enforcement Releases (AAERs) by the U.S. Securities and Exchange Commission. While correctly estimating accrual manipulation can be difficult, the use of AAERs (as in Dechow, Hutton, Kim, and Sloan (2012)) provides an alternative measure to test the possible substitution among different manipulation techniques. The tests report a higher likelihood of AAERs by covered firms when adoption of alternative data increases.

In the satellite coverage setting, I find similar results on REM. While overproduction does not appear to change significantly, treated firms exhibit higher cash flows from operations and stronger advertising spending. These results are accompanied by an increase in total accruals. Taken together, the findings in both settings support the intuition on the role of market-provided proxies in reducing real activity manipulation. Sophisticated investors respond negatively to indications of REM, which changes the behavior of managers of firms covered by the data. In both settings, I find some evidence on the substitution from real actions to other forms of accounting manipulation when the real actions become more observable.

This paper contributes to two main strands of literature. First, by investigating the effect of alternative data on managerial behavior and financial markets, it illuminates the role of new sources of trading signals on misreporting, contributing to the literature on the effects of Big Data on financial markets and firms (Zhu (2019); Katona et al. (2024); Farboodi, Matray, Veldkamp, and Venkateswaran (2021); Bonelli and Foucault (2023); Dichev and Qian (2022); Agarwal, Qian, and Zou (2021); Blankespoor, Hendricks, Piotroski, and Synn (2022)). The paper also contributes to this literature by providing a novel labor-based measure of the adoption of alternative data by sophisticated investors. By studying the impact of a continuously measured signal of firm performance, the results contrast with the literature on the adverse effects of frequent disclosure (Ernstberger et al. (2017); Kraft et al. (2018)), highlighting the importance of the source of information affecting managers' behavior.

Second, the paper contributes to the literature on REM (Kothari et al. (2016); Roychowdhury (2006)) and on the substitution between different combinations of inefficient actions and misreporting (Irani and Oesch (2016); Terry, Whited, and Zakolyukina (2022)). By providing new evidence on the potential substitution from real actions to accrual-based misreporting, which regulatory efforts are more focused on counteracting, the paper contributes to the literature on the effectiveness of new signals in enabling financial markets to discipline value destruction (acting as a private enforcer (Becker and Stigler (1974); Landes and Posner (1975)).

The rest of the paper is structured as follows: Section 2 develops the research question and the testable hypotheses. Section 3 describes the empirical strategy and the data used. Section 4 reports the results and the discussion. Section 5 concludes.

2 Hypotheses Development and Conceptual Framework

Alternative data offer direct insights into a company's economic activities, which are not directly influenced via its accounting system.[3] I investigate whether the increased availability of this kind of data influences managers' behavior, particularly in their decisions to act inefficiently and misreport, by enhancing external monitoring.

While this paper focuses on the effects of alternative data on real activity manipulation, previous findings highlight the usefulness of these signals in observing real-time performance of firms. Transaction information provides consistent and accurate signals of consumers' demand and is associated with returns predictability, as shown by Agarwal et al. (2021). Alternative data sources also provide incremental information with respect to GAAP revenues, although analysts do not fully incorporate this information in their forecasts,[4] as documented in Dichev and Qian (2022).

The introduction of Big Data has also influenced managers' disclosure decisions. Froot, Namho, Gideon, and Ronnie (2017) examine how real-time corporate sales data affect managers' disclosure of private information and find that managers tend to under-disclose so they can trade opportunistically. Blankespoor et al. (2022) using a large database of credit card transactions, show that managers disclose abnormally low sales early in the quarter, often due to litigation risk concerns. Zhu (2019) demonstrates that managers

[3]Alternative data contain information useful for predicting firms' performance (as earnings or revenues) that is obtained from sources other than financial statements or analysts' forecasts. For a detailed description, see Deloitte for Financial Services (2016).

[4]The incomplete incorporation of the alternative data in analysts' reports is relevant for this paper, as it highlights the importance of these sources of information, but it still allows an informational advantage for sophisticated investors using the data.

respond by reducing opportunistic trading, due to the decrease in their informational advantage.

While financial markets have become more informative over time (Bai, Philippon, and Savov (2016)), the overall effects of the new sources of information remain ambiguous. Farboodi et al. (2021) provide evidence of strong cross-sectional differences on the effect of Big Data on firms, with the documented increase in price efficiency concentrated among large and growth stocks.[5] Dessaint et al. (2024) find that alternative data have increased forecast accuracy over the short term, at the cost of a decrease in accuracy over longer horizons.

The availability of new data allows sophisticated investors to trade on more precise signals of firms' current performance, as shown by Bonelli and Foucault (2023). With the introduction and adoption of alternative data, sophisticated investors can observe a (potentially imprecise) signal of firms' performance in real time, including ahead of the announcements of financial results. While the coverage of alternative data is not randomly assigned, firms do not choose whether to be covered by a vendor. The coverage appears to be determined by the availability and existence of the data.

The introduction of alternative data reduces the information asymmetry between managers and investors regarding the firm's current performance. Given the internal information

[5] Additionally, Begenau, Farboodi, and Veldkamp (2018) show how the effects at the firm level have mostly benefited large firms, which, through a lower cost of capital, have been able to grow larger over time. Farboodi et al. (2021) highlight the cross-sectional differences in the effect of data availability on firms, noting that the rise in aggregate price efficiency is mostly driven by growth stocks. Yet the introduction of new sources of trading signals can also present drawbacks, both for firms and for traders without access to the data (Katona et al. (2024))), in contrast to what has been found by Zhu (2019).

systems available within public companies, there is little reason to believe that managers themselves acquire new insights about their own firms' performance from these external data sources. However, it is impossible to exclude the possibility that managers might learn about their competitors' operations through the use of alternative data.[6]

To examine how the introduction of alternative data can affect managerial short-termism, I rely on the intuition from the model of Stein (1989). Myopic actions, such as aggressively discounting goods to boost short-term revenues or cutting R&D spending to inflate near-term earnings, can be interpreted as borrowing from future performance (see also Terry (2023)). If alternative data provide a clearer signal of firm performance, investors may be better equipped to detect such manipulation, as they can observe a less biased indicator of cash flows and earnings. Additionally, the availability of a more precise indicator of true realized performance can reduce the incentive to engage in myopic actions, as their perceived benefits are likely to diminish when it becomes harder to mislead investors.

While some forms of myopia, such as unusually low cash flows resulting from aggressive discounting, are more directly observable through alternative data, others, like underinvestment in R&D, are only indirectly affected. Specifically, a signal of current performance may not directly influence a manager's decision to invest in R&D. However, if the signal impedes myopic actions from deceiving the market, it may in turn lower the perceived benefits of manipulation and thus reduce its prevalence. Whether these effects will be observable in the data is an empirical question.

[6] See Further Discussion & Alternative Channels section for a more detailed discussion of this issue.

Upon detecting signs of real activity manipulation, sophisticated investors are likely to become skeptical of the firm's prospects. This skepticism can lead to adverse consequences for the company's stock for two nonmutually exclusive reasons. First, myopic actions can harm future profitability (Kothari et al. (2016); Cohen and Zarowin (2010)).[7] Second, strategies like aggressive discounting may indicate underlying poor performance. Negative stock performance increases the cost associated with pursuing myopic actions. Consequently, the increased observability from the new data can lead to a reduction in the perceived value of real activity manipulation.[8] The chosen measure to capture external stakeholders' monitoring of firms' performance is short selling.

Short selling is an appropriate measure of monitoring as it is costly: investors must incur expenses to maintain their positions, which requires some degree of confidence in their future outlook. Previous work highlights the role of short selling in deterring both accrual-based earnings management (Massa et al. (2015)), and real earnings management (Cai et al. (2024)). If sophisticated investors can detect manipulation through alternative data, this should influence their trading, including increased short selling. For this effect to be present empirically, investors must use the new data to trade, with their activity reflecting an increased monitoring of real activity manipulation.

The relation between the frequency of performance observation and managerial behavior remains ambiguous. Evidence on the effects of increased disclosure frequency (Gigler

[7]That REM damages future performance is not a consensus result, with Gunny (2010) finding that firms that engage in REM tend to see positive subsequent performance, casting doubt on the opportunistic nature of the action.

[8]As in Fischer and Stocken (2004), the effectiveness of manipulation depends on the information available to the market. I investigate whether alternative data inform speculators about REM, implicitly reducing the effectiveness of REM and thus its benefits.

et al. (2014); Ernstberger et al. (2017); Kraft et al. (2018)) shows a worsening of managerial myopia. Since more frequent disclosure can lead to more real activity manipulation, this paper's focus on real action, together with the timing of the information release, provides an immediate comparison with the mechanism studied here. I investigate whether alternative data can mitigate myopic actions or, as with more frequent disclosure, whether increased market pressure will exacerbate these operational issues. While more frequent disclosure can accentuate managers' short-termism, alternative data provides insights into the operations of the firm.

The distinction in the origin of information, that is, from third-party observations of real transactions rather than firm-provided disclosures, may lead to a different effect on real activity manipulation. However, these new sources of information are not uniformly informative. As an example, being able to observe real-time sales might not provide any insight into the firm's investment in research and development but could illuminate aggressive discounting or overproduction. I hypothesize that sophisticated investors' ability to observe a firm's operations can lead to different outcomes than those caused by more frequent disclosure.

Empirically, to study the disciplinary mechanism discussed and the potential effects on managers' behavior, I first test whether the introduction of alternative data, along with their adoption by sophisticated investors, leads to a measurable change in trading. If the new data are informative about REM, stronger shorting of a firm showing signs of real activity manipulation should be observed. The first hypothesis of this paper states:

> **Hypothesis 1:** Following the introduction and adoption of alternative data useful for forecasting revenue, a more intense negative reaction from financial markets to a firm engaging in REM can be observed.

The hypothesis requires that financial markets will react *more negatively* to results that can be attributed to REM for firms that are covered by the data. It underscores the distinction between the availability and adoption of alternative data. The introduction may have no effect if investors do not respond. While it is impossible to directly observe the informational inputs used by sophisticated investors, who are often secretive to preserve their informational edge, I measure adoption indirectly by examining investments in Big Data-related human capital.[9] Hypothesis 1 represents the first stage of the mechanism studied. If managers face stronger shorting when manipulating their operations, the cost of engaging in REM will increase.

The main effect investigated in the paper is whether the new data will lead to a decrease in managerial short-termism by enhancing external monitoring. Previous survey (Graham et al. (2005)) and experimental (Cade et al. (2024)) findings indicate that real activity manipulation might be common in US public firms. Cade et al. (2024) reports that close to 30% of firms have manipulated real activities manipulation.[10] If the new signals reveal inefficient actions, the perceived value of meeting and beating targets through these actions

[9]Specifically, I use data on employees' skills at securities firms to construct a proxy for Big Data adoption. This measure captures the extent to which firms are investing in and using alternative data in their trading strategies. The construction of the measure is discussed in the following section.

[10]The forms of manipulation studied in this paper, consistent with the definition used by Cade et al. (2024), are based on the insights of Roychowdhury (2006): overproduction to lower the cost of goods sold, aggressive discounting to boost short-term revenues, and abnormal cuts to discretionary spending. These practices do not violate GAAP nor prompt an investigation by the SEC. However, they can be costly in terms of future performance.

may decrease. The following hypothesis is based on the insight that managers choose their means of manipulation according to the relative (perceived) costs of each action.

Hypothesis 2: More intensive external monitoring, induced by alternative data, is expected to lower managers' propensity to manipulate real activities.

Whether more intense monitoring by investors will inevitably reduce real activity manipulation is unclear. Samuels, Taylor, and Verrecchia (2021) show that the relation between monitoring and accounting manipulation is not monotonic. Higher levels of monitoring can induce stronger manipulation through the effects on earnings relevance to investors. However, alternative data are available only to a subset of market participants, which may reduce the concern about strengthening managers' incentive to manipulate through a higher value relevance of earnings.

Even if increased monitoring does lead to a reduction in some inefficient actions, the overall effect of the increased information available on firm performance is unclear. When managers are pressured by short-term incentives to achieve financial goals, they weigh the capital market benefits of success against the cost of manipulation (Bird, Karolyi, and Ruchti (2019)). A manager whose implicit cost of real activity manipulation has increased can respond in several ways. First, overall manipulation might decrease (Li and Liu (2022)). Alternatively, the manager might shift from observable real actions to others for which the new data are uninformative. Finally, managers might become more aggressive with their revenue recognition or accounting manipulation (AEM). The equilibrium outcome depends on external factors, as shown by Irani and Oesch (2016); Cohen, Dey, and Lys (2008); Ewert and Wagenhofer (2005). Which outcome will be realized is an empirical question investigated here. The hypothesized reduction in real activity manipulation leaves the effect on other forms of earnings management ambiguous.

The introduction of the data could raise some concerns over changes to the equilibrium that are beyond those investigated here. The importance given to earnings by investors might be affected, thus changing the incentives of the manager to pursue different types of misreporting. However, only a fraction of investors observes and can process the signal from alternative data, while much of the market relies on traditional information sources. This distinction between agents who observe the alternative data mitigates the concerns related to the change in value relevance of earnings, with the associated changes in incentives for the managers of covered firms. Although alternative data are informative for investors, managers maintain an informational advantage over the overall performance of the firm, such as soft information used in reporting where managerial discretion is involved.

To summarize, I first test whether the introduction and adoption of alternative data lead to a measurable change in investors' monitoring of firms' performance. I then investigate whether the increased monitoring, by raising the cost of manipulation, reduces real activity manipulation among firms covered by the data.

3 Empirical Strategy & Data

To investigate the effects of alternative data on the ability of stakeholders to monitor and consequently on managerial behavior, I rely on two distinct settings. The main setting is based on transaction data from point-of-sale collection. I then repeat the analysis using satellite data covering retailers. The use of two settings allows me to examine the effects of alternative data on monitoring and misreporting as a broader phenomenon, rather than one limited to a specific data source.

Transaction Data Analysis

Measuring the coverage of point-of-sale transaction data presents some challenges. As more sources of alternative data have become available, tracking which firms have been covered has become more complicated. Several vendors have released credit card and other point-of-sale data, while new store-level information, such as scanner data, has also emerged. Given the difficulty of measuring which firms have been covered by alternative data and when, I rely on a different approach: I measure the adoption of Big Data by sophisticated investors. By examining how and when sophisticated investors use alternative data, I can infer the extent of coverage and its impact, sidestepping the direct measurement issues associated with tracking individual transaction data sources and firms.

Using detailed data provided by Revelio Labs, I measure different dimensions of skills employed by sophisticated investors over time. I have manually selected, from among the recorded possibilities, a list of skills and job titles likely related to Big Data (such as “software engineer" and “business data scientist") employed by firms active in securities trading. This selection process helps identify these firms’ adoption of Big Data capabilities, providing insight into their use of alternative data. This measure proxies for the adoption of human capital related to alternative data by sophisticated investors. This novel proxy can be observed over a long period and enables the measurement along two dimensions:

total staff related to alternative data use in a given year and its ratio to the total number of employees in the industry. I use both measures to provide a comprehensive understanding of how the adoption and use of Big Data evolves among sophisticated investors. Figure 1 reports the number of profiles related to the adoption of alternative data.

While my adoption measures provide insights into investors' use of alternative data, how I determine which companies the data cover requires additional discussion. I rely on a list of firms whose products are covered by a data provider covering store-level transactions. This type of data has been shown to be useful for inferring the performance of consumer product manufacturers (Dichev and Qian (2022)). The interaction of the adoption measure with this coverage indicator based on store-level transactions proxies for treatment intensity.

Testing the Informativeness of Alternative Data

To test whether the introduction of new sources of information will reduce real activity manipulation, it is necessary to first investigate the usefulness of alternative data in enhancing monitoring. Hypothesis 1 states that the introduction and adoption of alternative data will inform sophisticated investors about a firm's real performance before that firm announces its results, allowing investors to short the firm's stock. To investigate this, I test whether investors react to potential signs of REM before the earnings announcement.

For this, I rely on short positions data at the monthly level from IHS Markit to compute two variables that should capture the amount of shorting in a stock, the short interest index (SII) and utilization. SII is computed as the number of shares on loan divided by the total number of shares outstanding, while utilization employs the same numerator scaled by the total number of shares that are registered in inventory. The two measures capture similar dynamics, with higher values indicating more shorting.

While alternative data can be informative and used throughout the year, I focus on the weeks between the end of the fiscal year and the announcement of results. This period matters for two main reasons. First, previous findings show that yearly results are the most relevant ones to investors (Stubben (2010)). While deviations can be observed throughout the first three quarters, it is likely that investors will scrutinize firms most closely around the fiscal year-end. Second, the weeks following the end of the fiscal year occur after the realization of results but before their announcement. If alternative data do help investors infer a firm's performance, the more impactful time to do that would be ahead of announcement, as this period would convey the greatest informational advantage. This specific period is used to measure the shorting of firms covered by the data when their upcoming announced results indicate manipulation. To define abnormal deviation from expected behavior, I rely on an approach similar to that of Roychowdhury (2006) and compute the residuals from an industry-year regression of cash flows from operations, R&D expenditures, and overproduction on a set of predictors. These measures, although subject to concerns about model specification, should capture whether a firm is behaving abnormally in a given period. The data used to compute the measure are from yearly financial statements (Compustat).

Given my focus on alternative data in informing investors about firms' performance before results are publicly available, the measures are based on the announced financial information. If the data do help detect REM, the announced results should correlate with investors' trading in the weeks between the end of the fiscal year and the public announcement. A similar test is designed to study whether, during the same period, short positions relate to discretionary accruals[11] reported in the upcoming announcements.

The intuition for this analysis is that the measure of REM could be capturing something unrelated to the construct of interest.[12]

These tests are designed to investigate whether, in the pre-announcement period, traders take short positions in stocks of firms that have manipulated their real activities before the announcement of their financial results. The design, which consists of comparing the weeks between fiscal year-end and the earnings announcement before and after the introduction of the alternative data, tests the usefulness of these new signals for measuring real performance. A depiction of the short analysis timeline is provided in Figure 2.

In this section, I present the estimation testing Hypothesis 1 in the first setting: the adoption of transaction data. To test the hypotheses, the regression design is the following:

$$
\begin{aligned}
y_{i,j,t} = \alpha &+ \zeta_1 \text{coverage}_i + \zeta_2 \text{number of profiles}_T + \zeta_3 \text{coverage}_i \times \text{number of profiles}_T + \\
&+ \zeta_4 \text{coverage}_i \times \text{IND}_{i,j,T} + \zeta_5 \text{number of profiles}_T \times \text{IND}_{i,j,T} + \\
&+ \boldsymbol{\beta}\ \mathbf{coverage_i \times number\ of\ profiles_T \times IND_{i,j,T}} + \\
&+ \Gamma_{i,j,T} + \alpha_{j,T} + \alpha_t + (\Gamma_{i,j,T} + \alpha_{j,T} + \alpha_t) \times \text{IND}_{i,j,T} + \epsilon_{i,j,t}
\end{aligned}
\tag{1}
$$

[11]The measure of discretionary accruals is estimated following previous work by: Dechow et al. (1995); Jones (1991); Dechow et al. (1995), and Irani and Oesch (2016))

[12]This additional test based on discretionary accruals aims to verify that the results are due to financial markets being able to observe REM before the announcement of the results. If the REM measure captured other aspects of the firms' actions, results could arise based on information that is not actually available yet, such as accruals.

where $y_{i,j,t}$ is the outcome of interest (SII and utilization) computed for firm i, in industry j in the year-month t. This measure is regressed on the variable of interest ($\text{coverage}_i \times \text{number of profiles}_T \times \text{IND}_{i,j,T}$) and firm controls ($\Gamma_{i,j,T}$), along with industry-year fixed effects ($\alpha_{j,T}$), and a year-month fixed effects (α_t). The coefficient of interest in this regression is β, associated with $\text{coverage}_i \times \text{number of profiles}_T \times \text{IND}_{i,j,T}$, which should capture whether the measure of manipulation (IND) is associated with shorting during the weeks between the end of fiscal year and the earnings announcement for firms covered by the data (coverage), when sophisticated investors adopt the new sources of data (number of profiles).

Following the insights from Gipper, Leuz, and Maffett (2019), I interact the measure of manipulation (IND) with both the fixed effects and the controls. Given that the variable of interest is an interaction ($\text{coverage}_i \times \text{number of profiles}_T \times \text{IND}_{i,j,T}$), the control variables and the fixed effects should also be interacted with the potential confounding effect of IND. Using the same specification, the analysis is repeated for both the measures of REM and for discretionary accruals. Hypothesis 1 predicts that firms covered by the newly introduced sources of data will face stronger shorting when exhibiting signs of REM ($\beta > 0$).

The research design helps me isolate the effect of the new data from the rest of the information environment. I naturally cannot observe directly how sophisticated investors use alternative data sources. But anecdotal evidence suggests that alternative data is advertised to investors as a tool to gain an informational advantage ahead of earnings announcements. Thus it is reasonable to focus on the time window I select (between the fiscal year-end and the earnings announcement) to test the informativeness of the data.

The main threat to the identification is that the estimated coefficient of interest β could be capturing something unrelated to the treatment or the measure of manipulation. To address this concern, I provide a placebo test in each setting by first dropping the treated firms and then choosing a random sample of firms that are not included in the definition of treatment. By repeating the same estimation described above, I test whether the results could be driven by the design or other confounding factors, instead of the mechanism of interest.

To further corroborate the analysis, I investigate whether the shorting of firms covered by the data is potentially profitable, thus creating a monetary incentive to sophisticated investors to monitor and trade based on the information. By using a similar estimation to the one highlighted in equation (1), I study the effects of alternative data on returns during the post-announcement period of firms that show signs of real earnings management. Specifically, I test whether firms showing stronger signs of REM, which are included in the treatment group, experience more negative returns in the three months following their earnings announcements. The analysis serves two purposes: first, to assess whether increased monitoring leads to a negative stock performance and, second, to verify whether the increased informativeness of alternative data concerning REM translates into profitable trading strategies, such as shorting before the announcement of results.

The main limitation of the tests for Hypothesis 1 is their reliance on ex post measures of real activity manipulation derived from reported accounting numbers. Although placebo tests and return-based analyses help strengthen the empirical strategy, the approach remains constrained by the fact that only the presence of alternative data coverage can be observed, not the actual content of the data. Accessing this content is unfortunately infeasible.

A Further Investigation of Alternative Data Informativeness

To provide additional evidence on the usefulness of alternative data, I analyze detailed advertising data. This additional setting allows me to employ data where the content is directly observable at a granular level. Advertising is a key component of discretionary spending, providing a direct link between the measures studied in the main analysis. I use the setting to provide further evidence of the mechanism underlying Hypothesis 1 on the informativeness of alternative data.[13]

AdIntel (provided by Nielsen) is a large database of product ads across multiple channels (TV, newspapers, digital, etc.), measured at the occurrence level (Shapiro et al. (2021)). The data are collapsed at the date level, to measure daily advertising expenditures of firms. Brands, once assigned the parent company, are then matched to accounting and financial data (Compustat), stock prices (CRSP), and short positions (IHS Markit). The final product is a database of daily ad expenditures at the firm level, which allows me to quantify the contemporaneous effect of changes in firms' behavior on the stock market. Aggregating at the month level, I investigate whether a deviation from advertising patterns leads to stock market reactions. For this test, I use the entire calendar and fiscal year.

A measure of abnormal monthly advertising spending is computed based on the expected level of spending at the industry-month level, controlling for firm characteristics and

[13] There is a concern related to advertising expenditure due to the nature of the ad market, where most of the slots are auctioned weeks in advance. However, as noted by Shapiro et al. (2021) and Hristakeva and Mortimer (2021), firms can choose to buy last minute advertising space for a considerable premium from specialized agencies.

yearly industry performance. To alleviate concerns related to the use of an abnormal measure in the estimation, I repeat the analysis using the simple month ad spending of the firm. Both measures are used to measure whether the stock market reacts to changes in ad spending. Together with SII and utilization, I measure the effect of abnormal ad expenditures on volume and bid-ask spreads, as the changes (recorded in alternative data) might be measured only by sophisticated investors, which could have re-distributive effects among market participants.

The advertising setting allows me to directly measure the outcome of managers' choices, without having to rely on measures of manipulation based on financial statements. Using a measure of cost shielding (Bloomfield, Gipper, Kepler, and Tsui (2021)), the analysis of the relation between advertisement and stock market outcomes can be expanded to validate whether the monitoring reflects managerial incentives. When managers are remunerated on "top-line" income, thus benefiting from cost-shielding provisions, their compensation might not be affected by an increase in advertising expenditure, which would otherwise lower the income measure used for determining performance. If the data enhance monitoring, an increase in effort (ad spending) for a manager benefiting from cost-shielding should not lead to the same stock market outcomes as for those compensation based on net income. This prediction is aimed at testing whether a possible positive stock market reaction to an increase in advertising expenditures reflects the monitoring channel discussed in the rest of the paper.

Measuring the Effects of Alternative Data on Misreporting and Real Activity Manipulation

Whether the monitoring role of alternative data will be reflected in managerial behavior is the core question investigated in this paper. The main outcomes of interest are the measures of real activity manipulation described by Roychowdhury (2006) and Irani and Oesch (2016): overproduction (to lower COGS), abnormally low cash-flows from operations (due to aggressive discounting to boost short-term revenues), and cuts to discretionary expenses (such as R&D and advertisement). To investigate the potential spillover effects of a reduction of real activity manipulation, I also study the effects of alternative data both on the amount of total accruals and on other measures of accounting manipulation, AAERs and restatements.[14]

The estimation approach follows the insight of Chen, Hribar, and Melessa (2018) and does not rely on a two-stage estimation approach with residuals as a dependent variable. In the one-stage estimation, I include the relevant predictors of the "normal" level of the variable of interest (following Roychowdhury (2006) and Irani and Oesch (2016)) and a series of additional controls based on the literature (size, return on assets, market-to-book ratio, leverage, and institutional ownership). To test the firms' response in REM following the

[14]The data on the SEC AAERs are provided by USC Marshall (see Dechow et al. (2012)). The data on restatements are provided by Audit Analytics. Following Choudhary, Merkley, and Schipper (2021), I identify Big-R restatements as one of the outcomes of interest.

introduction of the data, I estimate the following:

$$
\begin{aligned}
REM_{i,j,T} = &\alpha + \Omega_{i,j,T} + \zeta_1 coverage_i + \zeta_2 \text{number of profiles}_T + \\
&+ \Omega_{i,j,T} \times \text{number of profiles}_T + \\
&+ \boldsymbol{\beta} \textbf{coverage}_i \times \textbf{number of profiles}_T \\
&+ \Psi_{i,j,T} + \alpha_{j,T} + (\Psi_{i,j,T} + \alpha_{j,T}) \times \text{number of profiles}_T + \epsilon_{i,j,t}
\end{aligned}
\tag{2}
$$

where $REM_{i,j,T}$ denotes the measure of REM of interest and $\Omega_{i,j,T}$ the predictors for the normal level described in the literature (Roychowdhury (2006); Irani and Oesch (2016)). Based on the results of Chen et al. (2018), the measure of REM is regressed on the relevant predictors of the construct of interest together with the variable of interest in a single stage. $\Psi_{i,j,T}$ denotes a series of firm-level controls, and $\alpha_{j,T}$ is the industry-year fixed effects. The coefficient of interest is β, associated with the adoption of alternative data by sophisticated investors interacted with the coverage of the data. Greater adoption should be associated with less REM for firms covered by the proxies ($\beta < 0$ for overproduction, $\beta > 0$ for R&D expenditures and cash-flows from operations), given the heightened monitoring described in Hypothesis 1. To avoid identification concerns, given that the construct of interest is an interacted variable (coverage$\times$number of profiles), I interact the level of adoption with the predictors of the normal level of the outcome, the firm controls, and the fixed effects.

As in the analysis of short positions, I repeat the tests using the same placebo treatment sample. Keeping the observed measure of adoption, I rely on the estimation outlined in Eq. 2 using the placebo treatment. I use this test to provide additional evidence on the effect of transaction data on managers' behavior and to minimize the concerns that the findings could be attributed to the research design.

While my main focus is managers'n myopic actions, studying the impact of the same data sources on accrual-based earnings management can provide insights into the degree of substitution between real activity manipulation and aggressive accounting practices. A similar specification to that of REM is used to estimate the effect of alternative data on AEM:

$$\begin{aligned} accruals_{i,j,T} = & \alpha + \Gamma_{i,j,T} + \zeta_1 coverage_i + \zeta_2 \text{number of profiles}_T + \Psi_{i,j,T} + \alpha_{j,T} + \\ & + \Gamma_{i,j,T} \times \text{number of profiles}_T \\ & + \boldsymbol{\gamma} \textbf{coverage}_i \times \textbf{number of profiles}_T \\ & + \Psi_{i,j,t} + \alpha_{j,T} + (\Psi_{i,j,t} + \alpha_{j,T}) \times \text{number of profiles}_T + \epsilon_{i,j,t} \end{aligned} \tag{3}$$

where $accruals_{i,j,T}$ denotes the total accruals of firm i, in industry j, in year T. $\Gamma_{i,j,T}$ are the predictors of accruals based on the literature and the rest of the terms are described above. If the reduction in REM is accompanied by an increase in AEM, γ should be positive. Other outcomes related to accounting manipulation can also be used to provide additional insights into the effects of alternative data. For AAERs (restatements), I create an indicator variable taking the value 1 if a firm is subject to an enforcement action (Big-R, small-r) in a given year. This indicator variable is regressed on a set of controls and fixed effects, together with the variable of interest (similar to the above).

To better tie together the analyses discussed so far, I repeat the estimation of 2 using the measure of monthly advertising expenditure as the outcome variable. This allows me to better triangulate the findings and reduce reliance on measures of real earnings management derived from financial statements.

Satellite Data Analysis: Additional Evidence and Robustness

The second type of alternative data I study is satellite imagery. In 2011, vendors began offering measurements of the parking lot usage of US retailers. Sophisticated traders have since acquired this information to assess the volume of customers visiting stores. While the data only approximate retailers' profitability, they can be mapped into revenues, giving investors insight about customer trends before firms release their results and conduct earnings calls. This early insight can help investors anticipate financial outcomes and make better trades. The introduction of satellite data as a treatment has been used in the literature (see Zhu (2019); Katona et al. (2024), and Bonelli and Foucault (2023)), which shows its usefulness and which is further affirmed by its widespread adoption.

Satellite imagery coverage of retailers is observed through data provided by RS Metrics. This data provider offers records of covered retailers, enabling the identification of treated and control firms within the sector. Table S1 provides the summary statistics between the covered firms and the rest of the retailers available in Compustat.

Using satellite data as a secondary setting for robustness purposes allows me to examine whether the findings from the transaction data analysis generalize beyond the main setting. In particular, by employing a setting with different treatment and control groups, I test whether the monitoring channel and its impact on managerial behavior persist across different types of alternative data. I examine a shorter timeframe around the introduction of satellite data (2009–2013) to reduce potential confounding effects from the introduction of other data sources. Given the nature of the data and the clearly identifiable introduction, I employ a traditional difference-in-differences estimation approach within industries, without using the measure of adoption of alternative data by sophisticated investors. Further details of the identification strategy are provided in the appendix.

The use of satellite data does present some challenges. First, the signal extracted from these data is coarser, as satellite imagery is primarily informative about foot traffic. Additionally, the measures examined in this paper are often better suited for manufacturers. For example, while it is possible to construct a proxy for overproduction in the retail sector, its interpretation requires more caveats and assumptions.[15] Finally, the total number of firms (both treated and those used as controls) is lower than in the main setting. This smaller sample reduces the power of the tests. To partially address the limited sample size, I follow the approach of Roychowdhury (2006) and focus on the "suspicious region" (defined as firms reporting net income just above zero) as the one that is most likely to exhibit signs of real activity manipulation.

[15]The overproduction measure, based on cost of goods sold and inventory changes, may capture similar behavior for both retailers and manufacturers, particularly if retailers benefit from volume discounts from distributors that reduce their COGS.

4 Results & Discussion

The Informativeness of Alternative Data

According to Hypothesis 1, financial markets will react negatively to signs of REM when firms are covered by alternative data. The first set of results presented here examines whether the stock market responds to the introduction of alternative data. To distinguish between the effects of the information disclosed in financial statements and provided by alternative data, I focus the analysis on short positions during the period between the end of the fiscal year and the announcement of the results. If alternative data illuminate firms' performance, the data will provide the strongest informational advantage during this period.[16]

The first setting is based on the adoption of alternative data by sophisticated investors (number of profiles) and the coverage of firms provided by a vendor. Table 2 reports the results on the effects of the adoption of alternative data on capital markets when firms exhibit signs of manipulation. The outcome of interest is the Short Interest Index (SII) in the period following the end of the fiscal year. This measure reflects investors' short positions in a stock, with higher values indicating stronger shorting. I rely on two measures of real activity manipulation, following Irani and Oesch (2016), *RM* and *RM overproduction*.[17]

[16]The lag between the end of the fiscal period, when the performance is realized, and the announcement of the results is particularly well suited to study changes in behavior of sophisticated investors. While it is hard to observe *when* each source of alternative data is released to investors, the length of the period of interest should allow the signal to be observed. Additionally, during this period, financial results are realized, but they are not public information, as the announcement has not yet been made, increasing the likely informativeness of the new sources of information studied here.

[17]The measures are calculated as follows: *RM = -abnormally low CFO - abnormally low DISEXP* and *RM overproduction = overproduction*. The appendix provides details on the definition of the measures and their

Under the hypothesis investigated, higher REM should lead to stronger shorting for firms covered by the alternative data.

The results of Table 2 indicate that treated firms exhibiting higher deviation from their expected behavior (real activity manipulation) do face more intense shorting before their performance is announced when the adoption of new data is higher. An increase in the standard deviation of the measure of real earnings management is associated with an increase in SSI of approximately 5% for treated firms. The results are similar for both REM measures (columns (1) and (3)) and are robust to controlling for other performance variables, such as *ROA* and *net income* (columns (2) and (4)).[18] Using *discretionary accruals* as a measure of AEM, I repeat the analysis in the same setting. The results (columns (5) and (6)) do not indicate any sizable negative reaction by financial markets to possible accruals manipulation before the announcement of results. The difference in outcomes can indicate two non-mutually exclusive explanations. First, the effects found for real activity manipulation may most likely not be driven by other related factors, such as performance. Second, the results related to REM, but not AEM, may be attributable to the specific managerial behaviors that are more directly observable through the proxies examined. However, since the measures of manipulation may be subject to measurement error, it is possible (although not likely) that the difference in results reflects variation in

estimation.

[18]Table A10 provides the results of additional analyses based on controlling for the EPS surprise (*suescore*) and an indicator taking the value of 1 if the surprise is negative. The results resemble those discussed in the main body of the paper, with greater real activity manipulation being associated with more shorting for firms covered by the data, even when controlling for the EPS surprise. These results are reported as additional evidence. However, they should be interpreted with caution. EPS forecasts can incorporate information extracted by the alternative data (as shown by Dichev and Qian (2022)). Thus the surprise measures commonly used in the literature would be directly affected by the treatment definition.

the quality of the proxies rather than true differences in behavior.

To further corroborate the results discussed, I provide two additional analyses: a placebo test and an analysis of the post-announcement returns. The placebo test, using the adoption of alternative data setting, is designed with a "treatment group" built using a random sample of firms not covered by the proxy studied. Table A11 provides the results,[19] which do not indicate any sizable reaction to REM. These findings provide further evidence that the results from the two settings are indeed driven by the data studied, which enhance monitoring. Additionally, the placebo test results help address concerns related to the estimation strategy, as they indicate that the observed effects (in Table 2) are driven by the treatment rather than by underlying time trends in the adoption measure (which is the same in both tests).

To test whether the increased shorting is potentially profitable, I investigate whether REM is associated with lower returns during the months following the announcement of the results. Table A6 reports the effects of the adoption[20] of alternative data by sophisticated investors on covered stocks' returns. The three months following the announcement of the results is the period of interest. Using a similar specification to the test of short positions, I find that more real activity manipulation is associated with lower returns in the post-announcement period for firms covered by the new data.

[19]The placebo sample is built by considering a random selection of firms that are not covered by the transaction data coverage that I observe. The placebo sample is kept fixed in the analysis discussed later in the paper on the main effects of data on REM.

[20]Additional tables, available upon request, provide similar results in the relation between returns and alternative data induced monitoring by focusing on the growth rate of adoption. Using the growth rate captures a more contemporaneous and salient dimension of the adoption of alternative data by sophisticated investors. The results using this modified definition of adoption are stronger.

Taken together, the results discussed so far support Hypothesis 1. When alternative data are employed by sophisticated investors, the benefits of real activity manipulation shrink.

Further Evidence on the Monitoring Mechanism

To mitigate the potential concerns related to the use of ex post measures of manipulation in testing the monitoring mechanism, I rely on detailed micro-data on advertising, a type of alternative data, to investigate whether changes in advertising expenditures are reflected in the trading of sophisticated investors The motivation for using ad data is twofold. First, marketing expenses represent a component of discretionary expenditures, thus providing an immediate connection to the previous analysis. Second, observing advertising expenditures in a timely fashion allows for a more thorough investigation of the relationship between alternative data and monitoring. By testing whether contemporaneous changes in advertising are reflected in sophisticated investors' trading, this section provides additional evidence on monitoring by sophisticated investors.

I derive a measure of abnormal advertising at the month-industry level based on predicted expenditures aimed at capturing deviation in advertising efforts by the firm.[21] These tests resemble those of Hypothesis 1 but benefit from studying contemporaneous changes in advertising rather than a measure of REM computed ex post. Table 3 reports that an increase in abnormal advertising expenditures by a million dollars during a month reduces the short interest index (utilization) by 0.04% (0.08%), implying that a standard deviation increase in advertising leads to a 1.6% decrease in SII. I repeat the analysis using the simple monthly amount spent on ads by firms in the sample and find very similar results (see

[21] This measure is validated by using a raw measure of monthly advertising, with similar results.

Table A15), mitigating potential concerns related to a measure of abnormal expenditure. This additional robustness test indicates that the results are not driven by the choice of using a residual-based measure as the variable of interest.

For both measures of ad expenditures, I find that changes in expenditures lead to a decrease in volume and an increase in bid-ask spread, which is interpreted as a change in the distribution of information among market participants. These results complement those of Hypothesis 1 by providing an additional setting in which alternative data coverage informs sophisticated investors about firms' performance. Contemporaneous drops in advertising, implying abnormally low discretionary expenditures, lead to a negative financial market reaction.

In light of evidence on the average negative return on investment in advertising (Shapiro et al. (2021)), the findings on financial market reactions to declines in advertising spending should be interpreted with some caution. Investors may react to drops in advertising either because they view the cut as a sign of financial constraints, in line with reductions in discretionary spending often associated with real activity manipulation, or because they interpret it as a signal that managers expect weaker future demand, which highlights the informational content of advertising decisions. If investors believed that firms were overinvesting in advertising, a reduction in ad spending would be expected to trigger a positive market reaction. However, this is not supported by the empirical evidence.

To expand the discussion of the results based on advertising data, I repeat the analysis by more explicitly considering managers' incentive. Using an indicator taking the value of 1 if a manager's compensation contract includes cost shielding,[22] I investigate whether changes in advertising expenditures affect financial markets when managers bear fewer consequences of an increase in costs. Table 4 shows how the results on the relation between advertising expenditures and sophisticated investors are attenuated when managers benefit from cost-shielding.[23] Advertising expenses reduce shorting more when the manager bears the consequences of increasing the firm's costs. These results suggest that the relationship between discretionary expenditures and investors' reactions can depend on the manager's incentives. If managers are remunerated based on revenues, their choices regarding advertising might be insignificant to sophisticated investors.

The Effects of Alternative Data on Misreporting and Real Activity Manipulation

Having shown the informativeness of transaction data for sophisticated investors, I test whether this monitoring is reflected in changes in managers' behavior related to real and accounting manipulation. A reduction in real activity manipulation will be reflected in one or all of the following: less overproduction, higher discretionary expenditures, or higher cash flows from operations. Additionally, I investigate whether a potential reduction of

[22]As defined by Bloomfield et al. (2021), cost shielding is the practice of including compensation contract clauses that exclude some pre-specified costs. If a manager's compensation includes cost shielding, that person is remunerated based on higher-line results.

[23]The results of Table 4 indicate that financial markets react less to changes in advertising expenditures when managerial compensation is tied to top-line income. Additional discussion on the interpretation of the coefficients is provided in the table caption. Table A17 presents results similar to those in Table 4 but focuses on testing whether the coefficient on "Abnormal Advertisement Expenditures" differs significantly between firms with and without cost shielding. The findings indicate that financial markets respond differently to changes in advertising expenditures, depending on whether managerial compensation is shielded from such costs, and this difference is statistically significant. These results provide further evidence that financial markets react to increases in advertising expenditures primarily when the spending imposes a cost on managers.

REM is associated with an increase in likelihood of accruals manipulation.[24]

To test Hypothesis 2, which predicts that following more intense scrutiny by investors, managers will be less likely to pursue real activity manipulation, I estimate Equation 2 in the in the transaction data setting. The REM studied is composed of overproduction, changes in R&D expenditures,[25] and cash flow from operations.[26] All estimations are performed in a single stage, to avoid the concerns of Chen et al. (2018) of using a residual-based variable as the outcome.

Table 5 reports the results of the effects of transaction data adoption by financial markets on the measures of manipulation. The treated group is defined by the coverage of the proxy used, which provides a list of firms active in consumer staples and discretionary goods.[27] Firms covered by the data pursue less REM when adoption is higher, across the different measures. Specifically, covered firms exhibit higher R&D expenditures (column (3), corresponding to a 4% increase), higher cash flows from operations (column (4), equal to an increase by approximately 3%), and less overproduction (column (2)). The result on overproduction, while is directionally in line with the prediction of Hypothesis 2 is not statistically significant, and it has also a low economic magnitude, corresponding to a decrease of 0.7% in the measure. While it is always challenging to attribute the lack of results

[24]Despite not having a formal prediction for accruals manipulation, as discussed in the previous sections, the possibility of substitution between REM and AEM represents an interesting outcome. An increase in accruals manipulation could be reflected in higher total accruals.

[25]Using R&D expenditures instead of the more general *discretionary expenditures* allows for an easier comparison of the results with those from the literature on short-termism.

[26]All variables are defined in the appendix.

[27]The control group, used in Table 5, is defined by using the most similar firms to the covered ones in terms of TNIC (see Hoberg and Phillips (2010)). An alternative identification is based on using as a definition of treatment the union of the coverage and the five most similar firms, the results are available upon request.

to a specific factor, I propose two potential explanations for why overproduction appears less responsive than other measures. First, that this form of real activity manipulation may be less prevalent, although this interpretation is somewhat at odds with the findings on short positions. Alternatively, the overproduction measure may be less precise in capturing manipulation than the other proxies.

These results are accompanied by a larger use of total accruals (column (1), with an economic magnitude of an increase of approximately 1.5%) compared to the predicted normal level, possibly suggesting more aggressive accounting practices. The results on R&D expenditures highlight the different outcome induced by alternative data compared to more frequent disclosure. Stronger monitoring is associated with higher investment in R&D, contrary to the reduction reported in the literature on more frequent mandatory disclosure.

To further investigate the possible substitution from REM to AEM, I rely on data on AAERs, along with information on restatements provided by Audit Analytics. Given the difficulty in estimating accruals manipulation, studying the regulatory consequences of accounting manipulations provides additional evidence on the effects of alternative data adoption.[28] I define an indicator variable taking the value of 1 if the firm is subject to an AAER action (or a restatement) in a given year. Table 6 reports the findings of the analysis: greater adoption of alternative data is associated with a higher likelihood of facing an AAER but not a more general restatement for treated firms. While the magnitudes of these

[28] A shortcoming of using this type of data is the low frequency of the events considered. I further restrict the limited AAER sample to those actions related to revenues, inventory, COGS, and receivables.

coefficients are small, the evidence hints at the potential substitution from REM to AEM. The results related to AAERs, with the necessary caveat due to the low power of the test when using a strict treatment definition, may indicate that firms that reduce their real activity manipulation pursue more aggressive accounting. Taken together, the higher use of accruals and the higher likelihood of facing an enforcement action or a restatement seem to indicate that the adoption of alternative data, while reducing REM, might increase other types of misreporting.

To further corroborate the findings, I provide an additional placebo analysis to validate the empirical strategy and to investigate whether the findings are induced by something other than the treatment. Using the same placebo-treatment sample as the one used in Hypothesis 1, with the placebo treatment group drawn from the untreated firms, I repeat the analysis discussed in Table 5. The results (reported in Table A12) do not indicate any sizable effect, which strengthens the findings of the effects of alternative data adoption on manipulation.

Finally, to provide additional evidence on the role of alternative data monitoring in influencing the behavior of managers of covered firms, I conduct an analysis that combines the coverage of the proxy for revenues (based on transaction data) with the content of advertising data. Table 7 presents the results on advertising expenditures, using an identification strategy similar to that employed in the main analysis (based on coverage and REM). The findings show that firms covered by alternative data increase their investment in advertising as the adoption of the data by the financial industry rises. These results are robust across different specifications, clustering methods, and when using abnormal advertising expenditures as the outcome variable. Overall, this analysis integrates the different strands of evidence discussed in the paper and helps address concerns related to the use of financial statement-derived measures of real activity manipulation.

The results of the tests discussed above provide sufficient evidence for Hypothesis 2 by demonstrating that the introduction and adoption of alternative data leads to less real activity manipulation and a potential increase in aggressive accounting. These findings indicate the effectiveness of alternative data in reducing managerial myopia. Following the introduction and adoption of alternative data by sophisticated investors, covered firms exhibit less real activity manipulation. These results highlight the difference between alternative data and frequent disclosure, with the latter often inducing stronger short-termism. The differing outcomes can be understood by considering on the nature of the information disclosed. Alternative data provides insights into the actual performance of the firm, which is not directly influenced by managers' reporting incentives.

Satellite Data: Results & Discussion

For robustness, I replicate the analyses and test the two hypotheses using satellite imagery coverage as the source of treatment. This approach focuses on a specific sector (retail) and enables a sharp difference-in-differences estimation of the effects of interest. It also removes the need to rely on the human capital measure of alternative data adoption, providing an opportunity to better triangulate the earlier findings. However, this setting comes with two key limitations. First, the measures of real activity manipulation employed are typically better suited for manufacturers than for retailers. Second, the sample size is significantly smaller, which may affect statistical power.

Table S2 reports the results of the effects of satellite coverage on short positions, using the same outcomes and measures discussed above. While the results are less statistically significant, they resemble those found in the first setting. Retailers covered by satellite monitoring, during the post-period, face stronger shorting when exhibiting signs of REM. Similar results to those from the adoption setting hold also for *discretionary accruals*, showing that the negative financial market reaction is concentrated among the firms exhibiting signs of real activity manipulation. These results indicate that sophisticated investors react to deviations in firms' behavior when the new data are available and cover the firm. The results indicate that the effect of interest (higher shorting of covered firms showing signs of real activity manipulation) are not driven by the use of the human capital measure but by the treatment.

Overall, the results of the analysis support Hypothesis 1, similar to the adoption of transaction data. Despite the difference in statistical significance, both Table 2 and Table S2 provide evidence, in two distinct settings, of the monitoring role of sophisticated investors. The introduction and adoption of alternative data sources leads to a negative financial market reaction to signs of REM before the financial results are announced.[29]

Using satellite coverage as the second source of treatment, I investigate whether retailers whose performance is more observable have a similar response to the increase in monitoring recorded in the first setting. The empirical testing of Hypothesis 2 in the satellite setting

[29] Appendix B, available upon request, provides further empirical evidence of this mechanism, by studying the measures of manipulation (both REM and AEM) separately and with additional outcomes. Following Katona et al. (2024), I investigate whether the same mechanism that induces stronger shorting of firms exhibiting higher REM has sizable effects on the information environment of the stock. Specifically, I test whether there is an increase of bid-ask spread and a reduction of volume.

is complicated by two concerns: the smaller size of the treatment group and the weaker effects recorded in the first-stage analysis. To compensate for these issues, I rely on the insights of Roychowdhury (2006), and consider the firms' region of net income (scaled by lagged assets) between 0 and 0.05 as being firms most likely to exhibit REM, due to small loss avoidance. The coefficient of interest is the one associated to the indicator of this region interacted with the difference-in-differences estimator. As discussed in the previous sections, the predicted coefficient for accruals will depend on the trade-off between managerial incentives and substitution induced by the monitoring of REM. An increase in accruals, if accompanied by a decrease in inefficient actions, could indicate some substitution between REM and AEM. The outcomes of interest are total accruals, overproduction, advertising expenditures,[30] and cash flows from operations.

Table S3 provides the findings from the analysis of the effects of satellite coverage on retailers' actions. The satellite treatment, during the post-period, leads to slightly less overproduction (column (2), not statistically significant), higher investment in advertising (column (3)), and higher cash flows from operations (column (4)). The size of the effects in the "suspicious region" resemble those found in the transaction data setting. The results indicate an overall decrease in real activity manipulation across the different measures, accompanied by an increase in total accruals (column (1)).

The results discussed above should be interpreted with caution. While they support Hypothesis 2 in an alternative setting using a more traditional identification strategy,

[30]I rely on advertising expenditures in lieu of R&D as the industry studied (retailers) will most likely compete through marketing. The interpretation of the coefficient is the same, as advertising is one of the three components of *discretionary expenditures*.

they are concentrated in the "suspect region." This requires drawing on the insights of Roychowdhury (2006), who finds that firms reporting net income just above zero are most likely to engage in REM. Based on this rationale, it is reasonable to expect that the observed effect of alternative data on real activity manipulation would be most pronounced in precisely the area where this behavior is most likely to occur.

While the AAER and restatement analyses could not be used in the satellite setting, due to sample sizes, I provide additional evidence on managers' accounting choices becoming more aggressive following the decrease in REM. I study how the allowance for doubtful accounts changes, as a fraction of total accounts receivable, for treated firms. In principle, there is no expectation regarding a change on the estimation of doubtful accounts as a fraction of accounts receivable following the introduction of the data. In Table S4, I provide evidence of a negative association between doubtful accounts and (gross) accounts receivable for treated firms following the introduction of the satellite coverage.[31] This negative relation holds for several specifications, including the estimation of changes in doubtful accounts on changes in receivables. Along with the overall increase in accruals, the evidence on the allowance indicates a probable substitution from REM to more aggressive accounting.

Further Discussion & Alternative Channels

A significant issue related to the introduction of alternative data is whether managers of the firms covered by the data can also purchase and use it to gain insights into competitors' behavior. As discussed in previous sections, managers of firms included in alternative

[31] Additionally, the test provides a more intuitive representation: the share of the allowance of doubtful accounts receivable over total accounts receivable.

data sources may not gain new insights into their own firms' profitability or performance. However, they might be able to acquire valuable information about their competitor, such as the profitability of specific markets or customer demand for particular products. If so, they may alter their strategic decisions, even in the absence of any intent to manipulate.

While I do not have direct evidence on whether firms purchase the data discussed in this paper, this alternative mechanism seems less plausible. If managers were indeed reacting to enhanced information about their competitors, increased competitive behavior would be observed. This heightened competition could lead to market entry in specific segments, potentially explaining the rise in R&D investment reported in Table 5. However, increased competition would also be expected to depress prices, which, for a given level of sales, would result in lower cash flows from operations. This effect is not evident in the data, leading to doubts about the likelihood of this mechanism being the primary driver.

Another potential mechanism could help explain some of the observed results. In line with Kanodia and Mukherji (1996), the introduction of alternative data may serve as an external signal of a firm's cash flows, thereby reducing the information cost of investment. This reduction in information asymmetry could lead to an increase in investment, such as R&D and advertising, which might be mistakenly interpreted as a decline in myopic behavior. While this alternative explanation cannot be ruled out, I present evidence of a broad decrease in real activity manipulation (across various measures) and a potential rise in aggressive accounting. Taken together with the empirical findings related to short positions, these results suggest that the monitoring mechanism discussed is likely to drive the observed effects.

Given the use of AAERs in my analyses, it is necessary to consider whether the findings (increase in enforcement actions) reported in Table 6 stem from the rarity of these events. To mitigate these concerns, I study whether the adoption of alternative data leads to a higher probability of the firm being *investigated* by the SEC. I obtain raw data on all closed SEC investigations between January 1, 2000, and August 2, 2017, from Blackburne, Kepler, Quinn, and Taylor (2021). The database differs from the one used in the AAERs analysis, as it includes the nonpublic investigation by the SEC (with over 80% of the events not disclosed by the affected firms). These investigations "predict economically material declines in future firm performance" (Blackburne et al. (2021)).

Table A9 reports the results of a logit regression in which the outcome is an indicator taking the value of 1 if a firm is investigated by the SEC in any given month. The results indicate an increase, although statistically insignificant, in the likelihood of a firm being investigated being associated with the treatment definition or interaction of interest (treat · number of profiles). Although statistically insignificant, these results are informative, as they show that, when the sample expands, the effect remains directionally consistent with the findings from the AAERs analysis. While the larger sample provides more observations, it also introduces additional noise. In the AAERs sample, I focus specifically on cases involving accounting actions closely related to the behaviors under investigation (such as revenue recognition and COGS), whereas the broader analysis discussed here includes all SEC investigations, regardless of their nature. Overall, the results indicate that firms that are covered by alternative data might substitute from real activity manipulation to more aggressive accounting, leading to a higher likelihood of investigations and enforcement actions.

To better understand how the adoption of alternative data by sophisticated investors affects the information environment and the role of earnings announcements, I expand the analysis to investigate how markets react to unexpected earnings news. The results in Table A4 show that covered firms exhibit lower ERCs for a given level of negative earnings surprise ($Treat \cdot NumberofProfiles$). Specifically, in the case of negative EPS surprises ($Treat \cdot NumberofProfiles \cdot NegativeSurpriseIndicator$), covered firms experience smaller cumulative abnormal returns compared to firms not included in the data. The results with the triple interaction are statistically insignificant but can still be of interest.[32] Both for SUE score and negative surprise, the results indicate that firms covered by the data have lower earnings responses (on average) during the three-day event windows centered around the announcements.

If announced earnings become less relevant to financial markets, managers may face weaker incentives to meet or beat earnings targets. This could, in turn, reduce the motivation to manipulate real activities or to tailor manipulation strategies based on the content of alternative data. While it is impossible to directly observe how managers prioritize different performance targets, the evidence on increased monitoring, the decline in real activity manipulation, and the potential rise in aggressive accounting suggest that the monitoring mechanism proposed in this paper is likely the most relevant driver of the observed effects. If the introduction of new information sources reduces the relevance

[32] Although the measure of unexpected earnings and the definition of the treatment group differ, the results related to the SUE score have the opposite sign but a similar level of statistical significance, compared to those presented by Zhu (2019).

of traditional accounting announcements, the overall findings on the impact of data on managerial short-termism would still hold, though they would operate differently.

Whether the effects of the alternative data studied here will persist remains an open question. Asamoah, Massa, Mensah, and Tang (2025) show that the usefulness of alternative data declines as adoption becomes more widespread, with fund managers becoming less able to "extract alphas" from these signals. If sophisticated investors have less incentive to rely on a specific source of data for trading decisions, the effects documented here may not last, and the equilibrium could revert to one characterized by a greater reliance on real activity manipulation.

5 Conclusions

This paper studies the effects of the introduction and adoption of alternative data in disciplining real activity manipulation by informing sophisticated investors. I first show that the availability and adoption of alternative data is reflected in an increased monitoring of firms' performance ahead of earnings announcements. To measure the adoption of data by sophisticated investors, the paper introduces a novel measure of human capital investment in skills compatible with Big Data. Greater adoption leads to more intense monitoring for firms covered by the data, reflected in stronger shorting and lower returns for firms exhibiting signs of REM. Consistent with increased monitoring, I provide evidence for a decrease in managers' propensity to manipulate real activities (higher investment in R&D and higher cash-flows from operations).To provide further evidence on the mechanism, the analysis is repeated in a different setting: satellite coverage. Despite the lower statistical significance of the results, the evidence aligns with the findings of the adoption setting. Finally, the informativeness of alternative data in monitoring firms' performance is examined in the more granular setting of advertising. These findings validate the mechanism.

This paper provides evidence that new data sources do contribute to the monitoring of real activity manipulation. Managers respond by reducing REM, with some evidence of substitution toward AEM. Despite the use of various methodologies and the inclusion of placebo tests, the results should be interpreted with caution. The main results are based on noisy REM measures derived from financial statements.

With access to more content from alternative data, future research could more thoroughly investigate how these new data sources enhance monitoring. Additionally, access to granular information from these data sources would enable more precise measurement of managers' responses to increased monitoring, potentially clarifying the shift from real activity manipulation to aggressive accounting. As discussed in the advertising setting, the results on monitoring seem to strongly depend on the managers' incentives and compensation contracts. Further research could explore this mechanism in greater depth, emphasizing the importance of managerial incentives in shaping the output measures captured by alternative data.

6 Bibliography

References

Sumit Agarwal, Wunlan Qian, and Xin Zou. Disaggregated sales and stock returns. *Management Science*, 67:7167–7183, 2021.

Prince Elvis Asamoah, Massimo Massa, Albert Mensah, and Vicki Wei Tang. The early bird catches the worm: How lasting is the value of new, alternative data? *SSRN Working Paper*, 2025.

Jennie Bai, Thomas Philippon, and Alexi Savov. Have financial markets become more informative? *Journal of Financial Economics*, 122(3):625–654, 2016.

Gary S. Becker and George J. Stigler. Law enforcement, malfeasance, and compensation of enforcers. *The Journal of Legal Studies*, 3(1):1–18, 1974.

Juliane Begenau, Maryam Farboodi, and Laura Veldkamp. Big data in finance and the growth of large firms. *Journal of Monetary Economics*, 97:71–87, 2018.

Andrew Bird, Stephen A. Karolyi, and Thomas G. Ruchti. Understanding the "numbers game". *Journal of Accounting and Economics*, 68:2021–2061, 2019.

Terrence Blackburne, John D. Kepler, Phillip J. Quinn, and Daniel Taylor. Undisclosed sec investigations. *Management Science*, 67(6):3403–3418, 2021. doi: 10.1287/mnsc.2020.3805.

Elizabeth Blankespoor, Bradley E. Hendricks, Joseph Piotroski, and Christina Synn. Real-time revenue and firm disclosure. *The Review of Accounting Studies*, 27:1079–1116, 2022.

Matthew Bloomfield, Brandon Gipper, John D. Kepler, and David Tsui. Cost shielding in executive bonus plans. *Journal of Accounting and Economics*, 72(2):101428, 2021.

Maxime Bonelli and Thierry Foucault. Displaced by big data: Evidence from active fund managers. *SSRN Working Paper*, 2023.

N.L. Cade, J.L. Gunn, and A.J. Vandeberg. Measuring the Prevalence of Earnings Manipulations: A Novel Approach. *Journal of Accounting Research*, forthcoming, 2024.

Tianyu Cai, Lixiong Guo, and Yongxian Tan. Short seller monitoring and real earnings management. *Financial Review*, 59(1):203–225, 2024.

Wei Chen, Paul Hribar, and Samuel Melessa. Incorrect inferences when using residuals as dependent variables. *Journal of Accounting Research*, 2018.

Preeti Choudhary, Kenneth Merkley, and Katherine Schipper. Immaterial error corrections and financial reporting reliability. *Contemporary Accounting Research*, 38(4):2423–2460, 2021.

Daniel Cohen, Raj Mashruwala, and Tzachi Zach. The use of advertising activities to meet earnings benchmarks: evidence from monthly data. *Review of Accounting Studies*, 15, 2010.

Daniel A. Cohen and Paul Zarowin. Accrual-based and real earnings management activities around seasoned equity offerings. *Journal of Accounting and Economics*, 50(1):2–19, 2010.

Daniel A. Cohen, Aiyesha Dey, and Thomas Z. Lys. Real and accrual-based earnings management in the pre- and post-sarbanes-oxley periods. *The Accounting Review*, 83(3): 757–787, 2008.

P. Dechow, R. Sloan, and A Sweeney. Detecting earnings management. *The Accounting Review*, 70(2):193–225, 1995.

Patricia M. Dechow, Amy P. Hutton, Jung Hoon Kim, and Richard G. Sloan. Detecting earnings management: A new approach. *Journal of Accounting Research*, 50(2):275–334, 2012.

Olivier Dessaint, Thierry Foucault, and Laurent Fresard. Does alternative data improve financial forecasting? the horizon effect. *The Journal of Finance*, 79(3):2237–2287, 2024.

I.D. Dichev and J. Qian. The benefits of transaction-level data: The case of nielseniq scanner data. *Journal of Accounting and Economics*, 74(1):101495, 2022.

Jürgen Ernstberger, Benedikt Link, Michael Stich, and Oliver Vogler. The Real Effects of Mandatory Quarterly Reporting. *The Accounting Review*, 92(5):33–60, 09 2017. ISSN 0001-4826. doi: 10.2308/accr-51705.

Ralf Ewert and Alfred Wagenhofer. Economic effects of tightening accounting standards to restrict earnings management. *The Accounting Review*, 80(4):1101–1124, 2005. ISSN 00014826.

Maryam Farboodi, Adrien Matray, Laura Veldkamp, and Venky Venkateswaran. Where Has All the Data Gone? *The Review of Financial Studies*, 35(7):3101–3138, 10 2021.

Paul E. Fischer and Phillip C. Stocken. Effect of investor speculation on earnings management. *Journal of Accounting Research*, 42(5):843–870, 2004.

Deloitte Center for Financial Services. Alternative data for investment decisions: Today's innovation could be tormorrow's requirement. *Deloitte*, 2016.

Kenneth Froot, Kang Namho, Ozik Gideon, and Sadka Ronnie. What do measures of real-time corporate sales say about earnings surprises and post-announcement returns? *Journal of Financial Economics*, 125(1):143–162, 2017.

Frank Gigler, Chandra Kanodia, Haresh Sapra, and Raghu Venugopalan. How frequent financial reporting can cause managerial short-termism: An analysis of the costs and benefits of increasing reporting frequency. *Journal of Accounting Research*, 52(2):357–387, 2014.

Brandon Gipper, Christian Leuz, and Mark Maffett. Public Oversight and Reporting Credibility: Evidence from the PCAOB Audit Inspection Regime. *The Review of Financial Studies*, 33(10):4532–4579, 12 2019.

J.R. Graham, C.R. Harvey, and S. Rajgopal. The economic implications of corporate financial reporting. *Journal of Accounting and Economics*, page 3–73, 2005.

Katherine A. Gunny. The relation between earnings management using real activities manipulation and future performance: Evidence from meeting earnings benchmarks. *Contemporary Accounting Research*, 27(3):855–888, 2010.

Gerard Hoberg and Gordon Phillips. Product market synergies and competition in mergers and acquisitions: A text-based analysis. *The Review of Financial Studies*, 2010.

Sylvia Hristakeva and Julie H. Mortimer. Price dispersion and legacy discounts in the national television advertising market. *SSRN Working Paper*, 2021.

Rustom M. Irani and David Oesch. "analyst coverage and real earnings management: Quasi-experimental evidence". *The Journal of Financial and Quantitative Analysis*, 2016.

Jennifer J. Jones. Earnings management during import relief investigations. *Journal of Accounting Research*, 29(2):193–228, 1991. ISSN 00218456, 1475679X.

Chandra Kanodia and Arljit Mukherji. Real effects of separating investment and operating cash flows. *Review of Accounting Studies*, (1), 1996.

Zsolt Katona, Marcus O. Painter, Panos N. Patatoukas, and Jean Zeng. On the capital market consequences of big data: Evidence from outer space. *Journal of Financial and Quantitative Analysis*, page 1–29, 2024.

S.P. Kothari, Natalie Mizik, and Sugata Roychowdhury. "managing for the moment: The role of earnings management via real activities versus accruals in seo valuation". *The Accounting Review*, 91(2):559–586, 2016.

Arthur G. Kraft, Rahul Vashishtha, and Mohan Venkatachalam. Frequent Financial Reporting and Managerial Myopia. *The Accounting Review*, 93(2):249–275, 03 2018. ISSN 0001-4826. doi: 10.2308/accr-51838.

William M. Landes and Richard A. Posner. The private enforcement of law. *Journal of Legal Studies*, 4, 1975.

Qian Li and Shangqun Liu. Can third-party online sales disclosure help reduce earnings management? evidence from china. *Journal of Accounting and Public Policy*, 41(6):106977, 2022. ISSN 0278-4254.

Massimo Massa, Bohui Zhang, and Hong Zhang. The invisible hand of short selling: Does short selling discipline earnings management? *The Review of Financial Studies*, 28(6): 1701–1736, 2015.

Sugata Roychowdhury. Earnings management through real activities manipulation. *Journal of Accounting and Economics*, 42(3):335–370, 2006.

Delphine Samuels, Daniel J. Taylor, and Robert E. Verrecchia. The economics of misreporting and the role of public scrutiny. *Journal of Accounting and Economics*, 71(1):101340, 2021.

Bradley T. Shapiro, Günter J. Hitsch, and Anna E. Tuchman. Tv advertising effectiveness and profitability: Generalizable results from 288 brands. *Econometrica*, 89(4):1855–1879, 2021.

Jeremy C. Stein. Efficient Capital Markets, Inefficient Firms: A Model of Myopic Corporate Behavior. *The Quarterly Journal of Economics*, 104(4):655–669, 11 1989.

Stephen R. Stubben. Discretionary revenues as a measure of earnings management. *The Accounting Review*, 85(2):695–717, 2010.

Stephen J. Terry. The macro impact of short-termism. *Econometrica*, 91(5):1881–1912, 2023.

Stephen J Terry, Toni M Whited, and Anastasia A Zakolyukina. Information versus Investment*. *The Review of Financial Studies*, 36(3):1148–1191, 08 2022.

Amy Y. Zang. Evidence on the trade-off between real activities manipulation and accrual-based earnings management. *The Accounting Review*, 87(2):675–703, 2012.

Christina Zhu. Big Data as a Governance Mechanism. *The Review of Financial Studies*, 32(5): 2021–2061, 04 2019.

Summary Statistics and Variables Definitions:

Variable Dictionary

Variable	**Description**	**Frequency:**	**Source**	**Use**
Profiles	"profiles" denotes the measure of human capital employed by sophisticated investors related to big-data, based on job description. The measure is plotted in figure 1 and it is interpreted as a proxy for investment in big-data by sophisticated investors. The measure is interacted with an indicator for coverage, which denotes whether a firm is present in the database of store-level transactions (a reliable source of alternative data).	Yearly	Labor measure: Revelio Labs. Firms covered: hand collected coverage of database measuring store-level transactions.	Measure of interest
Profiles Growth	"profiles growth" denotes the growth rate of “profiles. This measure is meant to capture contemporaneous investment in big-data technologies. A lagged version is also used for some tests.	Yearly	Labor measure: Revelio Labs. Firms covered: hand collected coverage of database measuring store-level transactions.	Measure of interest
Satellite Coverage	Satellite coverage is an indicator variable taking value 1 if the retailer firm is covered by the data provider. The data contains the coverage (and the start of coverage) of all retailer firms whose parking lots were monitored by the data provider.	Yearly	RS Metrics	Measure of interest
Post	Indicator variable taking value 1 for the years following the introduction of the satellite coverage.	Yearly	RS Metrics	Time of adoption

Continued on next page

– continued from previous page

Variable	Description	Frequency	Source	Use
SII	Short Interest Index: = (quantity on loan/Number of Shares Outstanding) %. A higher value indicates a stronger shorting of the firm.	Monthly	Quantity on loan from IHS Markit, Shares outstanding: CRSP	Outcome
Utilisation	Utilisation = quantity on loan/(Stocks in inventory for lending)%. A higher value indicates a stronger shorting of the firm.	Monthly	IHS Markit	Outcome
B/A Spread (relative)	Bid-Ask spread (scaled by ask)	Monthly	CRSP	Outcome
Monthly Returns	Stock monthly return as reported in CRSP.	Monthly	CRSP	Outcome
Volume	Volume of Stock traded	Monthly	CRSP	Outcome
Monthly Spending (month_spend)	Raw total spending on advertisement by the firm.	Monthly	AdIntel	Measure of interest
Abnormal Spending (abn_spend)	Abnormal spending on advertisement by the firm. Computed as the residuals of a regression at the month-industry level of total spending.	Monthly	AdIntel	Measure of interest
AAER	Indicator variable taking value 1 if the firm has an AAER by the SEC in a given year. The enforcement actions considered are only those relate to revenue, receivables, COGS and inventory.	Yearly	USC Marshall AAER Dataset	Outcome
Restatement	Indicator variable taking value 1 if the firm has a restatement in a given year.	Yearly	Audit Analytics	Outcome

Continued on next page

– continued from previous page

Variable	Description	Frequency:	Source	Use
Big-R restatement	Indicator variable taking value 1 if the firm has a "Big-R" restatement in a given year. "Big-R" restatement are defined based on Choudhary et al. (2021), in which a restatement is considered Big-R if an 8-K filing accompanies the recognition.	Yearly	Audit Analytics	Outcome
Active Case	Indicator variable taking value 1 if the firm is subject to a SEC investigation in a given month.	Yearly	Blackburne et al. (2021)	Outcome.
Overproduction	Overproduction measure from (Roychowdhury (2006)), scaled by lagged assets.	Yearly	Compustat	Outcome
CFO	Cash-flows from operation (from Roychowdhury (2006)), scaled by lagged assets.	Yearly	Compustat	Outcome
Δ DisExp	Change in discretionary expenditures (from Roychowdhury (2006)), scaled by lagged assets	Yearly	Compustat	Outcome
Total Accruals	Total accruals scaled by lagged assets, based on Chen et al. (2018).	Yearly	Compustat	Outcome
Number of Stores	Number of physical retailer stores.	Yearly	Infogroup Database	Measure of interest
A/R over sales	Proportion of Account Receivables over total sales	Yearly	Compustat	Outcome
Net A/R over sales	Proportion of Account Receivables over total sales, net of allowance for uncollectible accounts.	Yearly	Compustat	Outcome
Ownership	Indicator variable taking value 1 for firms that are in the 90th percentile of institutional ownership for their stock.	Yearly	ISS	Control
RM	Measure of REM based on Irani and Oesch (2016) and Roychowdhury (2006): abnormally low cash flows from operations and abnormally low discretionary expenditures.	Yearly	Compustat	Measure of interest

Continued on next page

– continued from previous page

Variable	Description	Frequency	Source	Use
RM_ind	Indicator variable taking value 1 for firms that have measure of REM in the top-quartile of the distribution. The measure of REM is based on Irani and Oesch (2016) and Roychowdhury (2006): abnormally low cash flows from operations and abnormally low discretionary expenditures.	Yearly	Compustat	Measure of interest
RM_prod	Measure of REM-overproduction based on Irani and Oesch (2016) and Roychowdhury (2006): abnormally high production costs.	Yearly	Compustat	Measure of interest
RM_prod_ind	Indicator variable taking value 1 for firms that have measure of REM-overproduction in the top-quartile of the distribution. The measure of REM is based on Roychowdhury (2006).	Yearly	Compustat	Measure of interest
mdf_jones	Modified Jones measure of discretionary accruals. Used both in the discretionary form and the absolute value one. The measure is based on Irani and Oesch (2016) and Dechow et al. (1995).	Yearly	Compustat	Measure of interest
disc_ind	Indicator variable taking value 1 for firms that have measure of discretionary accruals in the top-quartile of the distribution. The measure of AEM is based on Irani and Oesch (2016) and Dechow et al. (1995).	Yearly	Compustat	Measure of interest
SUE score	Ratio of the absolute surprise to estimate dispersion for yearly EPS.	Yearly	I/B/E/S	Measure of interest

Continued on next page

– continued from previous page

Variable	Description	Frequency:	Source	Use
Neg. Surprise Ind.	Indicator variable taking value 1 if firm's announced results induce a negative suprise (SUE score < 1)	Yearly	I/B/E/S	Measure of interest
CAR	Cumulative Abnormal Returns computed as the sum across the three days event window (centered around announcement date) of the difference between the firm's daily returns and the market returns.	Yearly	CRSP	Outcome
Cost Shielding	Indicator variable taking value 1 for compensation contracts that include cost-shielding (see Bloomfield et al. (2021)) clauses.	Yearly	ISS	Measure of interest.

Adoption of big-data by sophisticated investors:

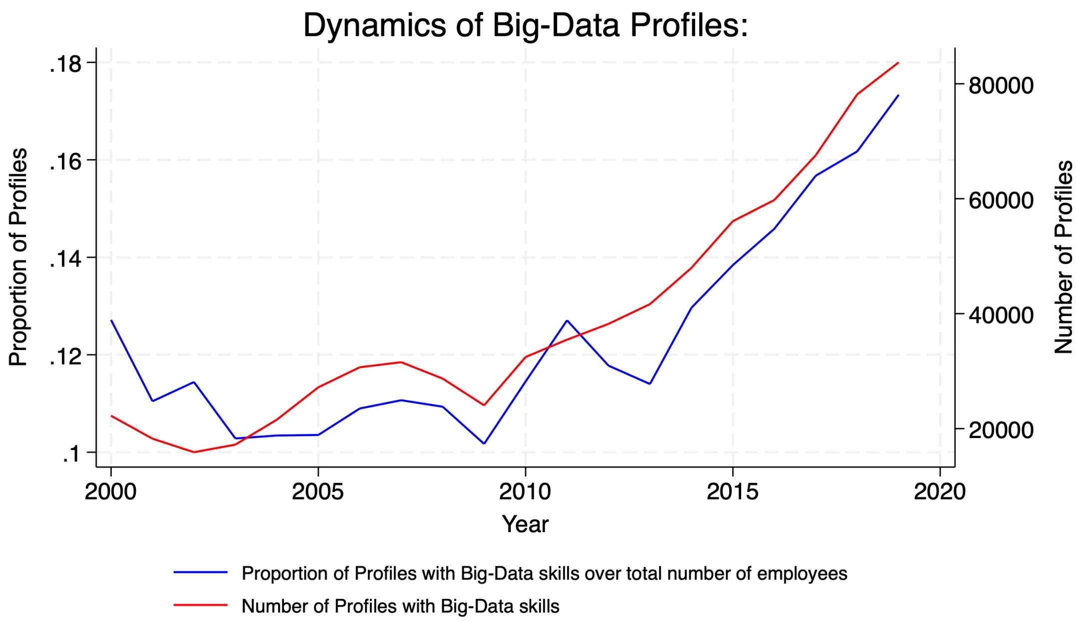

Figure 1. Time series of the number of employees with big-data skillset in securities-trading firms.

The figure reports the number of workers with skills compatible with big data employed by sophisticated investors. The measure captures the adoption of alternative data employed by securities trading firms. The source of the data is Revelio Labs, where all firms with NAICS code 523 (Securities, Commodity Contracts, and Other Financial Investments and Related Activities) active in the US were considered. The original data is based on the scraping of the universe of LinkedIn profiles, which allows to measure “backwards" the job description/title reported by the employees. Following the firms selection, I have downloaded all the work titles and manually selected the ones that are plausibly related to big-data. The measure depicted above (red line) is the count of employees of sophisticated investors whose skill-set is compatible with big-data. The blue line denotes the same measure normalized by the number of employees of the funds. The proxy captures the amount of investment in human capital related to alternative data. In addition to the count measure, I compute also the growth rate of it to capture contemporaneous changes of the interest of sophisticated inventors in the alternative data.

Short-position Analysis:

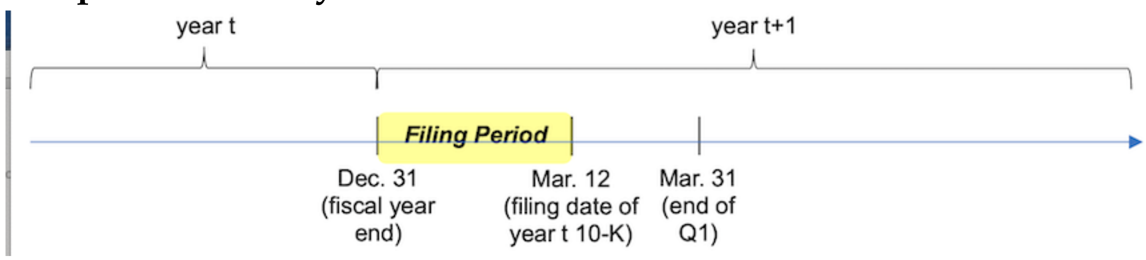

Figure 2. Graphical depiction of the timeline of short-position analysis

In the first set of tests for short positions, I investigate whether the measures of interest (SII and utilization) show a reaction from the introduction of the alternative data. To do so, I construct a measure of REM (based Roychowdhury (2006) and Irani and Oesch (2016)) and test whether companies covered (treated) by alternative data sources face more intense shorting when exhibiting higher likelihood of REM. Similarly, I repeat the analysis for discretionary accruals.

To test the usefulness of the data, I focus on the three months period following the end of fiscal year (yellow region) and I compare the effect on the outcome of interest of the introduction of the data (intuitively, I compare the yellow region before and after the introduction of the data for treated vs control firms). During this period the results are realized but the announcement has not taken place. I extend a similar analysis by focusing on the post-filing period, the three months period following the announcement of the results. For the post-announcement analysis, I investigate whether covered (treated) firms,exhibiting higher likelihood of REM, face lower returns.

Summary Statistics:

Table 1. Summary Statistics: Firms covered by Transaction Data

Market-to-Book Ratio

	25th percentile	Median	75th percentile	Mean	Standard Deviation
Non-covered Firms	1.265202	2.294769	4.127266	3.261018	8.529664
Covered Firms (treated)	1.531644	2.659261	4.55355	3.721518	8.037749
Observations	12883				

Size

	25th percentile	Median	75th percentile	Mean	Standard Deviation
Non-covered Firms	5.016094	6.606828	7.982368	6.438938	2.215915
Covered Firms (treated)	5.718658	7.592995	9.42337	7.386648	2.501288
Observations	12920				

The table reports summary statistics for two variables: market-to-book (MTB) ratio and firm size. A "Covered Firm" is defined as a firm included in the list of firms whose products are covered by the transaction data. A "Non-covered Firm" is not included in this list but is similar to a covered firm based on TNIC similarity. TNIC similarity, developed by Hoberg and Phillips (2010), measures the textual similarity of business descriptions in 10-K filings across public firms. For each firm in Compustat, the TNIC metric identifies the most similar firms based on these textual disclosures. This table compares firms known to be covered (i.e., included in the transaction data coverage list) with their most similar counterparts available in Compustat. Covered firms tend to be larger and exhibit higher MTB ratios. Table A1 presents the same statistics using an alternative treatment definition: covered firms and their most similar counterparts are compared to the broader universe of non-utility and non-financial firms in Compustat.

7 Tables

Transaction Data Setting

Short Analysis: Data Adoption and Capital Markets Reaction

Table 2. The Effects of Manipulation on Shorting Positions

	(1)	(2)	(3)	(4)	(5)	(6)
	SII	SII	SII	SII	SII	SII
Firm Covered (treated)	-0.62418	-0.59919	-0.62658	-0.51648	-2.11079	-1.90065
	(2.41712)	(2.44459)	(2.54751)	(2.56709)	(2.37081)	(2.39689)
Firm Covered (treated) × Number of profiles (adoption)	0.14061	0.15755	0.12610	0.12153	0.47764	0.49098
	(0.45904)	(0.46063)	(0.47255)	(0.47418)	(0.43943)	(0.44631)
Firm Covered (treated) × Number of profiles (adoption) × RM	0.04181***	0.04026***				
	(0.00375)	(0.00996)				
Firm Covered (treated) × Number of profiles (adoption) × RM overproduction			0.04532***	0.04608***		
			(0.00661)	(0.01246)		
Firm Covered (treated) × Number of profiles (adoption) × Discretionary Accruals					-0.10506*	-0.16357***
					(0.05452)	(0.05292)
Industry-Year F.E.	Yes	Yes	Yes	Yes	Yes	Yes
Year-Month F.E.	Yes	Yes	Yes	Yes	No	No
Performance Controls	No	Yes	No	Yes	Yes	Yes
Cluster	Industry	Industry	Industry	Industry	Industry	Industry
Obs.	74,752	74,752	74,752	74,752	72,777	72,777
Adj. R-Squared	0.190	0.195	0.187	0.193	0.138	0.148

Sample period: 2009-2018.

The table presents results on the effects of different measures of manipulation and misreporting (REM and AEM) on shorting activity. The period of interest is between the end of the fiscal year and the earnings announcement. Standard errors are clustered at the industry level and are provided in parentheses. *, ** , and *** indicate significance at the 10%, 5%, and 1% levels, respectively. The industry classification is defined at the two-digit SIC code. The treatment group is defined by firms covered by the alternative data source (transaction data), while the control group consists of the most similar firms based on textual similarity in filings (TNIC) that are not covered. The number of profiles refers to the total count of employees in securities trading firms that have job descriptions compatible with “big-data", used as a proxy for investment by sophisticated investors in alternative data. The outcome variable of interest is the short-interest index (SII). *RM* denotes the measure of real earnings management based on abnormally low discretionary expenses and cash flows from operations. *RM overproduction* is the measure of real earnings management based on overproduction. The controls include the market-to-book ratio, firm size, ROA, and scaled net income. Following the methodology of Gipper et al. (2019), I interact the measure of manipulation with both the controls and the fixed effects.

The table reports the results of an empirical test of Hypothesis 1, which states that firms covered by alternative data should face more negative reactions by financial markets when engaging in real activity manipulation compared to non-covered firms. The predicted sign of the coefficient of interest (the triple interaction in columns (1) to (4)) is: $\beta > 0$.
Higher levels of real activity manipulation (columns (1) and (3)), but not accrual-based manipulation (column (5)), are associated with increased shorting (in the weeks between fiscal-year end and the announcement of the results) of firms covered by the proxy when sophisticated investors have a high investment in human capital related to alternative data. The results also hold when controlling for the realized performance (columns (2) and (4)). Table A2 reports similar results by using a larger control group for the analysis. Table A3 reports similar results obtained without the interaction of the measure of manipulation with the control variables and the fixed effects, showing that the estimation choice of the interaction term for the measure of manipulation does not drive the main results.

.

Table 3. Abnormal Advertisement Expenditures and Stock Market Reaction

	(1) SII	(2) SII	(3) Utilisation	(4) Utilisation	(5) Volume	(6) Volume	(7) Bid-Ask Spread	(8) Bid-Ask Spread
Abnormal Ads Expenditure	-0.0444452***	-0.0220556***	-0.0785430***	-0.0655138***	-0.0089718***	-0.0051351*	0.0000104**	0.0000074*
	(-6.93)	(-2.82)	(-5.34)	(-3.56)	(-4.28)	(-1.88)	(2.37)	(1.77)
Lagged Abnormal Ads Expenditure		-0.0310855***		-0.0304113*		-0.0060668**		-0.0000017
		(-3.80)		(-1.69)		(-2.07)		(-0.39)
Industry-Month F.E.	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year-Month F.E.	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm F.E.	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Cluster	IndustryMonth	IndustryMonth	IndustryMonth	IndustryMonth	IndustryMonth	IndustryMonth	IndustryMonth	IndustryMonth
Obs.	12,500	11,626	13,059	12,137	12,552	11,675	14,988	13,943
Adj. R-Squared	0.590	0.597	0.669	0.678	0.514	0.511	0.484	0.551

Sample period: 2010-2019.

The table reports the results of the analysis on the effects of changes in advertisement expenditure patterns on capital markets. Abnormal Ads Expenditure measures deviation from normal behavior computed at the industry-month level based on expected monthly advertisement expenditure. Higher abnormal advertisement indicates ad spending above what is predicted for industry peers. The t-statistics are reported in parentheses (two-sided), standard errors are clustered at the industry-month level. *, ** , and *** indicate significance at the 10%, 5%, and 1% levels, respectively. Industry is defined using the two-digit SIC code. The coefficient of interest is the one associated with Abnormal Ads Expenditure.
Higher levels of advertisement are associated with contemporaneous lower shorting activity, as measured by SII and Utilization. The results indicate that an increase in advertisement leads to reduced shorting of the stock (columns (1)-(4)), accompanied by a decrease in volume and an increase in bid-ask spread (columns (5) to (8)). The results are interpreted as follows: deviations from expected behavior recorded through a granular source of advertisement expenditures (an alternative data source) are reflected in stock market reactions, thus lending credibility to the monitoring channel investigated in this paper. Table A16 reports a similar test with additional specifications for robustness, while Table A15 reports the equivalent analysis using raw monthly advertisement expenditure.

Table 4. Abnormal Advertisement Expenditures, Stock Market Reactions, and Cost Shielding

	(1) SII	(2) Utilisation
Abnormal Ads Expenditure	-0.030***	-0.001
	(-4.71)	(-0.46)
Cost Shielding Indicator	0.763***	0.229***
	(3.20)	(2.97)
Cost Shielding Indicator × Abnormal Ads Expenditure	0.004	-0.001
	(0.82)	(-0.76)
Industry-Month F.E.	Yes	Yes
Year-Month F.E.	Yes	Yes
Firm F.E.	Yes	Yes
Controls	Yes	Yes
Cluster	IndustryMonth	IndustryMonth
Obs.	6,352	6,358
Adj. R-Squared	0.622	0.624

Sample period: 2010-2019.

The table reports the results of the analysis on the effects of advertisement (as recorded in a source of alternative data) on capital markets. The t-statistics are reported in parentheses (two-sided), standard errors are clustered at the industry-month level. *, ** , and *** indicate significance at the 10%, 5%, and 1% levels, respectively. Abnormal Ads Expenditure is a measure of deviation from normal behavior computed at the industry-month level based on expected monthly advertisement expenditure. Industry is defined using the two-digit SIC code. *Cost Shielding Indicator* is a dummy variable taking the value of 1 if managers' compensation contracts include cost shielding provisions, defined as in Bloomfield et al. (2021).

Higher levels of advertisement are associated with contemporaneous lower shorting activity (SII, column (1)), but the results are attenuated when abnormal ads expenditure is interacted with the cost shielding indicator. The change in results suggests that capital markets might not consider advertisement expenditure increases as a valuable investment when managers' compensation is not affected negatively by the expenditure. The interpretation of the coefficient on abnormal advertisement expenditures in relation to shorting activity is as follows: for firms that employ cost shielding provisions, the total effect equals the sum of the coefficient on abnormal advertisement expenditures and the coefficient on the interaction between abnormal advertisement and the cost shielding dummy variable. See Table A17 for a more detailed discussion of the results.

Main Analysis: Data introduction and Effects on Misreporting

Table 5. The Effects of Alternative Data on REM & Accruals: Main Analysis

	(1) Total Accruals	(2) RM Prod	(3) Delta RD	(4) CFO Scaled
Firm Covered (treated)	-0.60712***	-0.05268	-0.01935***	-0.00374
	(-3.81)	(-1.06)	(-4.49)	(-0.17)
Firm Covered (treated) × Number of profiles	0.09686***	-0.00291	0.00615***	0.00831**
	(3.44)	(-0.78)	(3.09)	(2.24)
Industry-Year F.E.	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes
Cluster	Industry	Industry	Industry	Industry
Obs.	4731	9322	7473	9573
Adj. R-Squared	0.967	0.696	0.082	0.463

Sample period: 2009-2018.

The table reports the main analysis of the paper: the relation between alternative data coverage and manipulation. The t-statistics are reported in parentheses (two-sided), standard errors are clustered at the industry level. *, ** , and *** indicate significance at the 10%, 5%, and 1% levels, respectively. The treatment group is defined by firms covered by the alternative data source (transaction data), while the control group consists of the most similar firms based on textual similarity in filings (TNIC) that are not covered. Industry is defined at the two-digit SIC code. Number of profiles is the total count of employees in securities trading firms that have job descriptions compatible with “big-data", used as a proxy for investment by sophisticated investors in alternative data. The coefficient of interest is the interaction between coverage and adoption (number of profiles). All estimations are conducted using a one-stage approach, with the relevant predictors of "normal" levels (see Roychowdhury (2006)) included. Additional controls included in the estimation are market-to-book ratio, size, institutional ownership, *ROA*, and leverage. The table reports the results of an empirical test of Hypothesis 2, which predicts that firms that are covered by alternative data will reduce real activity manipulation due to the increased monitoring ability of sophisticated investors. The predicted signs, indicating a reduction in real activity manipulation, are: < 0 for *RM prod*, > 0 for Δ *R&D*, and > 0 for *CFO scaled*. The results indicate an overall decrease in real activity manipulation (columns (2) to (4)), with evidence of potential substitution towards more aggressive accounting practices, as shown in column (1).

Table 6. The Effects of Alternative Data on AAERs and Restatements

	(1) AAER Year	(2) AAER Year	(3) Restatement Year	(4) Restatement Year	(5) Big-R Year	(6) Big-R Year
Firm Covered (treated)	-0.01296***	-0.01648**	0.01927	0.02393	0.00400	0.00507
	(-2.74)	(-2.51)	(0.88)	(0.79)	(0.42)	(0.39)
Firm Covered (treated) × Number of profiles	0.00199**	0.00250**	-0.00231	-0.00307	-0.00029	-0.00030
	(2.50)	(2.26)	(-0.61)	(-0.63)	(-0.14)	(-0.12)
Model	LinProbMod	LinProbMod	LinProbMod	LinProbMod	LinProbMod	LinProbMod
Industry-Year F.E.	Yes	Yes	IndustryYear	Yes	Yes	Yes
Controls	No	Yes	Yes	Yes	Yes	Yes
Cluster	Industry	Industry	Industry	Industry	Industry	Industry
Obs.	14837	10753	14837	10753	14837	10753
R-Square	0.020	0.027	0.030	0.040	0.018	0.030

Sample period: 2009-2018.

The table reports the effects of alternative data coverage (treatment) on the likelihood of a firm facing an enforcement action by the SEC or a restatement. The t-statistics are reported in parentheses (two-sided), standard errors are clustered at the industry level. *, ** , and *** indicate significance at the 10%, 5%, and 1% levels, respectively. The treatment group is defined by the coverage of the proxy. The control is defined by the most similar firms in terms of TNIC similarity available in Compustat. Industry classification is defined at the two-digit SIC code. Number of profiles is the total count of employees in securities trading firms that have job descriptions compatible with "big-data", used as a proxy for investment by sophisticated investors in alternative data. The outcome variable is a dichotomous variable taking the value of 1 if a firm is the subject of an enforcement action by the SEC (AAER related to: revenues, inventory, accounts receivable, and COGS) or a restatement of interest in a given year. The results indicate that firms covered by alternative data have a higher likelihood of SEC enforcement actions. The findings are interpreted as additional evidence of the potential substitution from REM to AEM. Table A7 reports the equivalent results using a logit estimation. Table A8 reports a similar analysis considering using as control the rest of non-financial non-utilities companies available in Compustat. The results that are qualitatively similar.

Table 7. Advertisement Expenditures and Transaction Data Coverage

	(1) Month Expenditure	(2) Month Expenditure
profiles	-3.44976	-2.98072
	(-1.21)	(-0.93)
Firm Covered (treated)	0.40535	0.57387
	(0.29)	(0.26)
Firm Covered (treated) × profiles	1.53079***	0.81335**
	(5.81)	(2.23)
Industry-Month F.E.	Yes	Yes
Year-Month F.E.	Yes	Yes
Controls	No	Yes
Cluster	IndustryMonth	IndustryMonth
Obs.	16218	3859
Adj. R-Squared	0.183	0.378

The table reports the results of the analysis of how alternative data coverage affects advertisement investment, similar to the main analysis of the paper. The t-statistics are reported in parentheses (two-sided), standard errors are clustered at the industry-month level. *, ** , and *** indicate significance at the 10%, 5%, and 1% levels, respectively. Monthly Advertisement Expenditure is the total amount spent on ads by a firm in a given month, across its brands and different advertisement channels. Industry is defined at the two-digit SIC code. The treatment group is defined by the coverage of transaction data discussed in Tables 2 and 5. The control group is formed by the rest of public firms available in the advertisement database that are not included in the list of firms whose products are covered by transaction data. Number of profiles is the total count of employees in securities trading firms (sophisticated investors) that have job descriptions compatible with "big-data". The results are robust to different clustering of standard errors. "Controls" include the same controls used in the estimation of the "normal level" of discretionary expenditures and other industry-year level controls.
The table reports the results on advertisement expenditures (one of the components of discretionary expenditures used in REM analysis) using an estimation approach similar to that in Table 5, thus providing additional evidence on the previous results by being able to observe a more granular measure of discretionary expenditures. The coverage of alternative data increases the propensity of firms to invest in advertisement, implicitly reducing the propensity to engage in abnormal cuts to discretionary expenditures.

.

Satellite Coverage Setting

Testing the Informativeness of Satellite Data

Following the discussion of the first setting reported in the "Empirical Strategy & Data" section, I present the identification strategy using the satellite coverage setting. To measure the informativeness of satellite parking lot coverage for retailers, I estimate the following specification:

$$\begin{aligned} y_{i,j,t} = \alpha &+ \zeta_1 treat_i + \zeta_2 post_T + \zeta_3 treat_i \times post_T + \\ &+ \zeta_4 treat_i \times \mathrm{IND}_{i,j,T} + \zeta_5 post_T \times \mathrm{IND}_{i,j,T} + \\ &+ \boldsymbol{\beta}\ \mathbf{treat_i \times post_T \times IND_{i,j,T}} + \\ &+ \Gamma_{i,j,t} + \alpha_{j,T} + \alpha_t + (+\Gamma_{i,j,t} + \alpha_{j,T} + \alpha_t) \times \mathrm{IND}_{i,j,T} + \epsilon_{i,j,t}, \end{aligned}$$

where $y_{i,j,t}$ is the outcome of interest (SII and utilization) computed for firm i, in industry j in the year-month t. This measure is regressed on the variable of interest ($\mathrm{treat}_i \times \mathrm{post}_T \times \mathrm{IND}_{i,j,T}$) and firm-level controls ($\Gamma_{i,j,t}$), along with industry-year fixed effect ($\alpha_{j,T}$), and a year-month fixed effect (α_t). treat_i is a dichotomous variable taking value 1 for firms included in the list of covered retailers provided by RS Metrics and 0 for public firms active in retailing that are not covered. The coefficient of interest in this regression is β, associated with $\mathrm{treat}_i \times \mathrm{post}_T \times \mathrm{IND}_{i,j,T}$, which should capture whether the measure of manipulation (IND) is associated with shorting (during the filing period) for firms covered by the data (treat), following the introduction (post). Similar to Gipper et al. (2019), I interact the measure of manipulation (IND) with both the fixed effects and the controls used. Given that the variable of interest is an interaction ($treat_i \times post_T \times \mathrm{IND}_{i,j,T}$), the control variables and the fixed effects should also be interacted with the potential confounding effect of IND. Using the same specification, the analysis is repeated for both an measures of REM and for discretionary accruals. Hypothesis 1 predicts that firms that are covered by the newly introduced sources of data will face stronger shorting when exhibiting signs of REM ($\beta > 0$). All standard errors are clustered at the industry level. Given the small sample size of the analysis, which is limited to one sector (retail) and a relatively small treatment group, I use bootstrapped standard errors in the satellite coverage setting.

Similar to the adoption case, the main threat to the identification of the effects of data availability on monitoring would be that the estimated coefficient of interest β could be loading on something unrelated to the treatment or the measure of manipulation. To address this concern, I provide a placebo test by first dropping the treated firms and then choosing a random sample of firms that are not included in the definition of treatment. By repeating the same estimation described above, I test whether the results could be driven by the design or other confounding factors instead of the mechanism of interest.

Measuring the Effects of Satellite Data on Misreporting and Real Activity Manipulation

$$REM_{i,j,T} = \alpha + \Omega_{i,j,T} + \beta treat_i \times post_T + \Psi_{i,j,T} + \alpha_{j,T} + \epsilon_{i,j,t},$$

where $REM_{i,j,T}$ denotes the measure of REM of interest and $\Omega_{i,j,T}$ the predictors for normal level described in the literature (Roychowdhury (2006), Irani and Oesch (2016)). Based on the results of Chen et al. (2018), the measure of REM is regressed on the relevant predictors of the construct of interest together with the variable of interest in a single stage. $\Psi_{i,j,T}$ denotes a series of firm-level controls, and $\alpha_{j,T}$ is the industry-year fixed effect. The coefficient of interest is β, the one associated with the *treat*×*post*. The introduction of satellite data should be associated with a decrease of real activity manipulation for firms covered by the proxies ($\beta < 0$), given the heightened monitoring described in Hypothesis 1.

The satellite data introduction provides a sharp introduction of the treatment, which allows for the use of a difference-in-differences estimation to measure the impact of the new data on managers' behavior. The coefficient of interest is then the one associated with the difference-in-differences term $treat \times post$ with a predicted negative sign. This interaction can be complemented by using an indicator capturing the region of interest. The motivation behind the use of this indicator variable follows from Roychowdhury (2006), who argues that managers avoid small losses by managing real activities[33].

While the main focus of the paper is measuring the effect on REM, studying the impact of the same data sources on accrual-based earnings management AEM can provide insights into the degree of substitution between REM and AEM. A similar specification to the one describe above is used to estimate the effect of alternative data on AEM:

$$accruals_{i,j,T} = \alpha + \Psi_{i,j,T} + \zeta treat_i + \zeta_2 post_T + \boldsymbol{\gamma} treat_i \times post_T + \Upsilon_{i,j,T} + \alpha_{j,T} + \epsilon_{i,j,t},$$

where $accruals_{i,j,T}$ denotes the total accruals of firm i, in industry j, in year T. $\Psi_{i,j,T}$ are the predictors of accruals based on the literature and the rest of the terms are described above. If the reduction REM is accompanied by a potential increase in AEM, γ should be positive.

To further examine whether the introduction of satellite data influences managers' accounting choices, I provide an additional test focusing on the allowance for doubtful accounts receivables. In principle, an increase in monitoring does not have an effect on the recognition of what share of account receivables are doubtful. To investigate whether managers potentially change their accounting practices, I test whether the share of doubtful

[33]The regions used are those just above the zero net income and zero-net-income yearly growth rate, where the first one is motivated by small loss avoidance and the second by the practice of evaluating managerial performance using the prior year as a benchmark.

A/R (over total A/R) changes with the introduction of the data.

$$\frac{\text{Doubtful A/R}_{i,j,T}}{\text{Total A/R}_{i,j,T}} = \alpha + \kappa_1 treat_i + \kappa_2 post_T + \boldsymbol{\rho} treat_i \times post_T + \Upsilon_{i,j,T} + \alpha_{j,T} + \epsilon_{i,j,t},$$

A negative ρ would indicate that managers recognize a smaller fraction of A/R as potentially doubtful. An increase in monitor of overall performance of a retailer is not expected to either change the credit-worthiness of its client nor the estimation of what percentage of A/R could deteriorate. Together with an increase in total accruals, a change in recognition of allowance for doubtful A/R could indicate more aggressive accounting practices.

Advertisement Expenditures Measures:
The data on advertisement used in this paper is provided by AdIntel through Nielsen. The database used is based on the aggregation of occurrence-level records of advertisement expenditures aggregated to the brand-date level. From brands, I aggregate up to the firm-date level and finally match these data to information about the firm from Compustat. The level of observation used in the analysis is the month-firm one, where the date-level information is aggregated at the month-level. The "raw" measure of advertisement expenditure is defined as:

$$\text{Month Ads Expenditure}_{i,m} = \sum_{d} \text{ads expenditure}_{i,d}$$

where i denotes the firm identifier, d the daily date and m the month. Table A14 reports the summary statistics of this measure. From this month-level expenditure data, I derive a measure of "abnormal advertisement" as:

$$\text{Abnormal Ads Expenditure}_{i,m} = \text{Month Ads Expenditure}_{i,m} - \text{Month Ads }\hat{\text{E}}\text{xpenditure}_{i,m}$$

where "Month Ads $\hat{\text{E}}\text{xpenditure}_{i,m}$" is defined as the fitted value of a regression based on industry month predicted values of advertisement (accounting for both firms and industry specific predictors). The abnormal measure is meant to capture deviations from the expected value of advertisement.

Table 6 reports the summary statistics of the firms covered by the advertisement data compared to the rest of the Compustat sample used in other parts of this paper.

Table S1. Summary Statistics: Satellite Coverage

Market-to-Book Ratio

	25th percentile	Median	75th percentile	Mean	Standard Deviation
Non-covered Firms	.9110553	1.708529	3.192753	2.600041	8.623844
Covered Firms (treated)	1.972251	2.862233	4.623818	3.755892	3.480278
Observations	1531				

Size

	25th percentile	Median	75th percentile	Mean	Standard Deviation
Non-covered Firms	4.33033	5.908397	7.329075	5.775335	2.331987
Covered Firms (treated)	7.724507	8.786143	9.426671	8.535752	1.375301
Observations	1536				

The table reports the summary statistics for two variables: market-to-book ratio and size. A "Covered Firm" is a retailer included in the list of firms covered by satellite measurement. A "Non-Covered Firm" is a retailer for which satellite coverage is not available. Retail firms with satellite coverage have a higher market-to-book ratio and larger size.

.

Table S2. Satellite Data as Treatment: Effects of Manipulation on Shorting Positions

	(1) SII	(2) SII	(3) SII	(4) SII	(5) SII	(6) SII
Firm Covered (treated)	0.628	0.757**	0.550	0.471	1.468	1.834
	(1.20)	(2.24)	(0.81)	(1.45)	(1.00)	(1.13)
Firm Covered (treated) × Post Period	1.348***	1.359**	1.447***	1.734*	3.185	2.785
	(3.62)	(2.07)	(3.25)	(1.76)	(1.45)	(1.07)
Firm Covered (treated) × Post Period × RM	11.266*	17.743*				
	(1.94)	(1.86)				
Firm Covered (treated) × Post Period × RM overproduction			9.677***	13.176**		
			(3.21)	(2.34)		
Firm Covered (treated) × Post Period × Discretionary Accruals					25.576	26.950
					(1.20)	(0.81)
Industry-Year F.E.	No	Yes	No	Yes	Yes	Yes
Year-Month F.E.	No	Yes	No	Yes	Yes	Yes
Performance Controls	No	Yes	No	Yes	No	Yes
Cluster	Industry	Industry	Industry	Industry	Industry	Industry
Obs.	10,620	10,620	10,620	10,620	10,598	10,598

Sample period: 2009-2013.

The table reports the effects on shorting activity of different measures of manipulation and misreporting (REM and AEM) for retail firms. T-statistics are reported in parentheses. *, ** , and *** indicate significance at the 10%, 5%, and 1% levels, respectively. Bootstrapped standard errors are clustered at the industry level. The period of interest is between the end of the fiscal year and the announcement of the results. "Treated" refers to retail firms included in satellite coverage from 2011 onward. The control group is defined by the public retail firms that are not included in the satellite coverage. The industry classification is defined at the two-digit SIC code. The estimation approach for the satellite setting is described in the previous section. The outcome variable of interest is the short-interest index (SII). Higher values of SII are associated with stronger shorting of the stock. *RM* denotes the measure of real earnings management based on abnormally low discretionary expenses and cash-flows from operations. *RM overproduction* is the measure of real earnings management based on overproduction. The controls include the market-to-book ratio, size, ROA, and NI scaled.

The results indicate that retail firms covered by satellite data face stronger shorting when engaging in real activity manipulation, thus providing additional evidence for the monitoring mechanism.

.

Table S3. Satellite Data as Treatment: Effects on REM & Accruals

	(1) Total Accruals	(2) RM Prod	(3) Advertisement (scaled)	(4) CFO Scaled
Firm Covered (treated) × Post Period	0.029	0.011	-0.004	0.003
	(0.65)	(0.48)	(-0.68)	(0.60)
Firm Covered (treated) × Post Period × Suspect Region	0.583***	-0.031	0.020***	0.174**
	(5.84)	(-0.27)	(4.56)	(2.48)
Industry-Year F.E.	Yes	Yes	Yes	Yes
Firm F.E.	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes
Sample	2009-2013	2009-2013	2009-2013	2009-2013
Cluster	Industry	Industry	Industry	Industry
Obs.	803	1,101	808	1,154
Adj. R-Squared	0.758	0.977	0.902	0.829

Sample period: 2009-2013.

The table presents the main analysis of how alternative data (satellite coverage) affects retail firms' propensity to engage in real activity manipulation and misreporting. The t-statistics are reported in parentheses (two-sided). Bootstrapped standard errors are clustered at the industry level. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively. "Treated" refers to retail firms included in satellite coverage. The control group consists of retail firms in Compustat that are not covered. "Post" is an indicator variable equal to 1 if $year \geq 2011$, the beginning of satellite coverage data available for purchase by sophisticated investors. Industry classification is defined at the two-digit SIC code. *Suspect Region* denotes a firm having $\frac{NI_t}{A_{t-1}} \in [0, 0.005]$, which is associated with the highest likelihood of REM, following the argument of Roychowdhury (2006). The coefficient of interest is the interaction of coverage (treatment) in the post period, for firms with results falling in the suspect region. All estimations are run in a one-stage approach, with the relevant predictors of "normal" levels (see Roychowdhury (2006)) included. Additional controls included in the estimation are the market-to-book ratio, size, institutional ownership, *ROA*, and leverage. The table presents results from an empirical test of Hypothesis 2 using satellite coverage as the alternative data source. The predicted signs, indicating a reduction in real activity manipulation, are: < 0 for *RM prod*, > 0 for advertisement, and > 0 for *CFO scaled*. The results indicate a reduction in real activity manipulation and a potential substitution towards aggressive accounting practices.

Table S4. Satellite Data as Treatment: Effects on Doubtful A/R

	(1) Share of Doubtful A/R	(2) Doubtful A/R	(3) Doubtful A/R (scaled)	(4) Delta Doubtful A/R
Firm Covered (treated) × Post Period	-0.068*	9.337	-0.005	-0.001
	(-2.29)	(1.65)	(-1.31)	(-1.29)
Firm Covered (treated) × Post Period × A/R		-0.059***		
		(-9.98)		
Firm Covered (treated) × Post Period × A/R (scaled)			-0.028**	
			(-3.19)	
Firm Covered (treated) × Post Period × Delta A/R (scaled)				-0.300***
				(-6.13)
Industry-Year F.E.	Yes	Yes	Yes	Yes
Firm F.E.	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes
Sample	2009-2013	2009-2013	2009-2013	2009-2013
Cluster	Industry	Industry	Industry	Industry
Obs.	704	781	765	709
Adj. R-Squared	0.676	0.901	0.555	0.345

Sample period: 2009-2013.

The table presents an analysis of how alternative data (satellite coverage) affects accounting choices in firms covered by the data. Specifically, I test whether firms covered by satellite data exhibit different behavior in the post period in their accounting choices, measured as the share of doubtful accounts receivable over total accounts receivable. "Treated" refers to retail firms included in satellite coverage. The control group consists of retail firms in Compustat that are not covered. "Post" is an indicator variable equal to 1 if $year \geq 2011$, the beginning of satellite coverage data available for purchase by sophisticated investors. The t-statistics are reported in parentheses (two-sided), standard errors are clustered at the industry level. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively. Industry classification is defined at the two-digit SIC code. *A/R* denotes accounts receivable, *(scaled)* denotes the variable being scaled by lagged assets. Share of doubtful A/R is computed as the allowance for doubtful A/R over total (gross) A/R. The test examines whether the introduction of satellite coverage and the associated reduction in REM is accompanied by a change in accounting practices of treated retailers. The results indicate that retailers included in the satellite coverage exhibit changes in their allowance for doubtful accounts receivable. The results are interpreted as a more aggressive credit policy adopted by treated retailers.

.

Equations:

Real Earnings Management: Based on Roychowdhury (2006) and Irani and Oesch (2016):

- **Abnormal Cash Flows**: difference between actual cash flows from operations (scaled by lagged assets) and the predicted value from

$$\frac{CFO_{it}}{A_{i,t-1}} = \beta_1 \frac{1}{A_{i,t-1}} + \beta_2 \frac{sales_{i,t}}{A_{i,t-1}} + \beta_3 \frac{\Delta sales_{i,t}}{A_{i,t-1}} + \epsilon_{i,t}$$

  The resulting measure of REM is defined as RM_{CFO}. Higher level of REM will be associated with abnormally *low* cash-flows from operations.
  In the one-stage estimation I use $\frac{CFO_{it}}{A_{i,t-1}}$ and the same predictors as above in the estimation, together with other controls of interest.

- **Overproduction**: difference between realized production and predicted normal level. $PROD$ is defined as $PROD_{i,t} = COGS_{i,t} + \Delta INV_{i,t}$. The normal level is estimated based on:

$$\frac{PROD_{i,t}}{A_{i,t-1}} = \gamma_1 + \frac{1}{A_{i,t-1}} + \gamma_2 \frac{sales_{i,t}}{A_{i,t-1}} + \gamma_3 \frac{\Delta sales_{i,t}}{A_{i,t-1}} + \gamma_4 \frac{\Delta sales_{i,t-1}}{A_{i,t-1}} + \epsilon_{i,t}$$

  The resulting measure of REM is defined as RM_{PROD}. Higher level of REM will be associated with abnormally *high* production levels.
  In the one-stage estimation I use $\frac{\Delta PROD_{i,t}}{A_{i,t-1}}$ and the same predictors as above in the estimation, together with other controls of interest.

- **Discretionary expenditures**: $DisExp_{i,t} = R\&D_{i,t} + ADS_{i,t} + SG\&A_{i,t}$. Abnormal discretionary expenditures is estimated as:

$$\frac{\Delta DisExp_{i,t}}{A_{i,t-1}} = \phi_1 \frac{1}{A_{i,t-1}} + \phi_2 \frac{sales_{i,t-1}}{A_{i,t-1}} + \epsilon_{i,t}$$

  The resulting measure of REM is defined as RM_{DISEXP}. Higher level of REM will be associated with abnormally *low* discretionary expenditures.
  In the one-stage estimation I use $\frac{\Delta DisExp_{i,t}}{A_{i,t-1}}$ and the same predictors as above in the estimation, together with other controls of interest. When a component of discretionary expenditures is used, the same predictors as above are used.

To compute residuals based measures I require 15 observation in each industry-year pair (following Roychowdhury (2006)). RM_{IND} is an indicator taking value 1 if $RM_{IND} = -RM_{CFO} - RM_{DISEX}$ is in the top-quartile of the distribution. RM_prod_ind is an indicator equal to 1 if the measure of over-production is in the top-quartile of the distribution. All estimations are conducted at the industry-year level.

Discretionary accruals are estimated following Dechow et al. (1995). $discr\_acc\_ind$ is an indicator variable equal to 1 if the firm has discretionary accruals in the top-quartile of the distribution.

Appendix A:

Summary Statistics:

Table A1. Summary Statistics: Firms covered by Transaction Data (vs general control)

Market-to-Book Ratio

	25th percentile	Median	75th percentile	Mean	Standard Deviation
Non-covered Firm	.4628246	1.420422	3.195836	1.891235	11.49779
Covered Firm (treated)	1.303492	2.361863	4.241062	3.343124	8.445593
Observations	43769				

Size

	25th percentile	Median	75th percentile	Mean	Standard Deviation
Non-covered Firm	2.630872	4.319112	6.408292	4.512369	2.669828
Covered Firm (treated)	5.107315	6.770598	8.187628	6.608088	2.298213
Observations	43909				

The table reports summary statistics for two variables: market-to-book (MTB) ratio, firm size, and leverage. A “Covered Firm” is defined as a firm that is either directly included in the list of firms whose products are covered by the transaction data, or one of the five most similar firms in Compustat based on TNIC similarity. TNIC similarity, developed by Hoberg and Phillips (2010), measures the textual similarity of product descriptions in firms’ 10-K filings. “Non-covered Firms” comprise the remaining public firms in Compustat, excluding those in the financial and utility sectors. This comparison highlights systematic differences between covered and non-covered firms along key financial dimensions.

.

Adoption of Big-Data by Sophisticated Investors: Dynamics

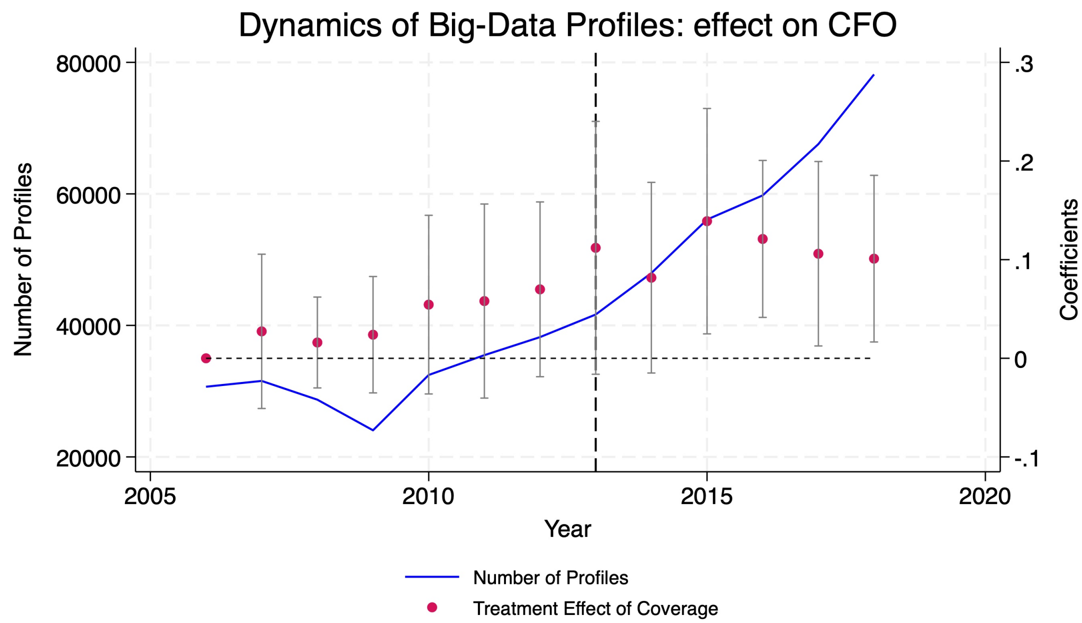

Figure 3. Dynamics of the Effects of Transaction Data on Cash-Flows from Operations.

The graph reports an estimation equivalent to equation 2, with year dummies interacted with the treatment. The treatment of interest is firm coverage by transaction data, and the outcome is cash flows from operations scaled by lagged assets (CFO_t/A_{t-1}). The coefficients (red dots) are reported together with their confidence intervals. A positive coefficient denotes a decrease in REM-CFO. The dashed vertical line denotes 2013, the first year following the foundation of Earnest Research (a prominent alternative data provider). The graph shows that firms did not react immediately to the introduction of the data to investors. There is no meaningful difference in the coefficients between covered and non-covered firms until the data was introduced.

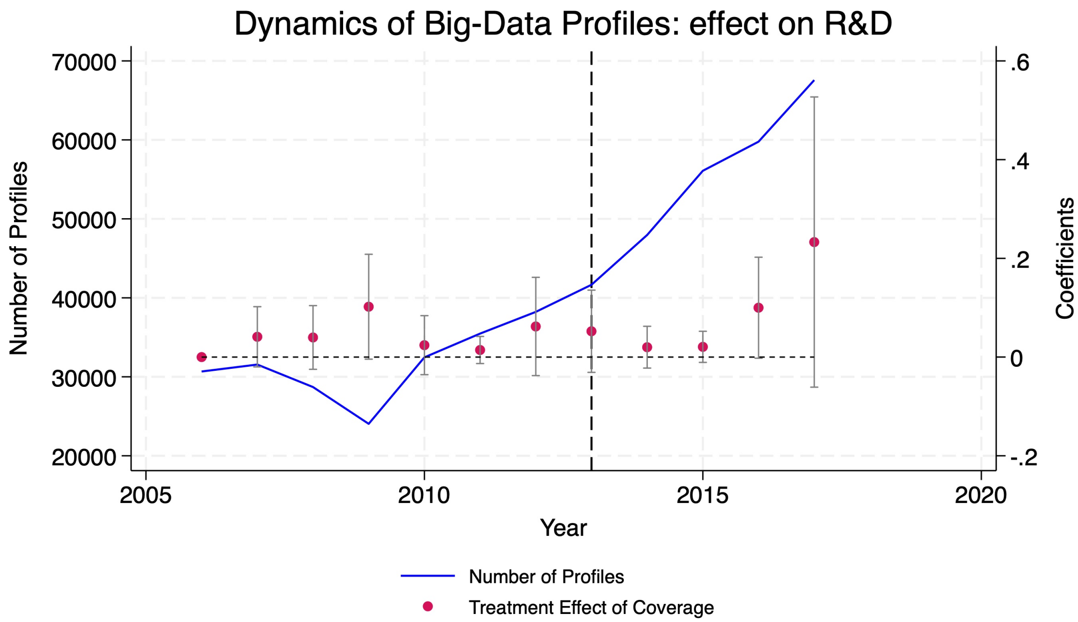

Figure 4. Dynamics of the Effects of Transaction Data on R&D Expenditures.

The graph presents an estimation equivalent to equation 2, with year dummies interacted with the treatment. The treatment of interest is the effect of firm coverage by transaction data, and the outcome is changes in R&D expenditures. The coefficients (red dots) are reported together with their confidence intervals. A positive coefficient indicates a decrease in myopic behavior, reflected in increased R&D investment. The dashed vertical line denotes 2013, the first year following the foundation of Earnest Research (a prominent alternative data provider). The graph shows that firms did not react immediately to the introduction of the data to investors. There is no meaningful difference in the coefficients between covered and non-covered firms until the data was introduced.

Short Position Analysis:

Table A2. The Effects of Manipulation on Shorting Position (General Control)

	(1)	(2)	(3)	(4)	(5)	(6)
	SII	SII	SII	SII	SII	SII
Firm Covered (treated)	-0.31743	-0.38920	-0.36524	-0.43688	-1.43427	-1.40285
	(1.76503)	(1.82673)	(1.75080)	(1.81416)	(1.74627)	(1.80955)
Firm Covered (treated) × Number of profiles (adoption)	0.10923	0.13455	0.10444	0.13016	0.30353	0.33367
	(0.33062)	(0.33669)	(0.32383)	(0.32821)	(0.33491)	(0.34874)
Firm Covered (treated) × Number of profiles (adoption) × RM	0.05568***	0.04600***				
	(0.01081)	(0.00967)				
Firm Covered (treated) × Number of profiles (adoption) × RM overproduction			0.05993***	0.05123***		
			(0.01089)	(0.01353)		
Firm Covered (treated) × Number of profiles (adoption) × Discretionary Accruals					-0.08864*	-0.12016**
					(0.04893)	(0.05098)
Industry-Year F.E.	Yes	Yes	Yes	Yes	Yes	Yes
Year-Month F.E.	Yes	Yes	Yes	Yes	No	No
Performance Controls	No	Yes	No	Yes	Yes	Yes
Cluster	Industry	Industry	Industry	Industry	Industry	Industry
Obs.	104855	104855	104855	104855	102735	102735
Adj. R-Squared	0.156	0.163	0.156	0.163	0.136	0.145

Sample period: 2009-2018.

The table reports the effects on shorting activity of different measures of manipulation and misreporting (REM and AEM). The period of interest is between the end of the fiscal year and the earnings announcement. The treatment group is defined by the coverage of the transaction data list of products and firms. The control group consists of all non-utility and non-financial firms available in Compustat. The t-statistics are reported in parentheses (two-sided), standard errors are clustered at the industry level. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively. The industry classification is defined at the two-digit SIC code. Number of profiles is the total count of employees in securities trading firms (sophisticated investors) that have job descriptions compatible with “big-data". The outcome variable of interest is the short-interest index (SII). *RM* denotes the measure of real earnings management based on abnormally low discretionary expenses and cash flows from operations. *RM overproduction* is the measure of real earnings management based on overproduction. The controls include the market-to-book ratio, size, ROA, and NI scaled. Higher levels of real activity manipulation, but not accrual-based manipulation, are associated with increased shorting in the weeks between fiscal year-end and the earnings announcement for firms covered by the proxy when sophisticated investors have a high investment in human capital related to alternative data. The results are similar to those presented in Table 2, indicating that the results discussed in the main analysis of this paper are not driven by the choice of control group.

Table A3. The Effects of Manipulation on Shorting Positions (No Interaction)

	(1) SII	(2) SII	(3) SII	(4) SII	(5) SII	(6) SII
Firm Covered (treated) × Number of profiles (adoption)	0.15210	0.15558	0.15210	0.15542	-0.01736	-0.01845
	(0.32)	(0.33)	(0.32)	(0.33)	(-0.04)	(-0.04)
Firm Covered (treated) × Number of profiles (adoption) × RM	0.00932*	0.01248***				
	(1.87)	(3.71)				
Firm Covered (treated) × Number of profiles (adoption) × RM overproduction			0.00802	0.01111**		
			(1.22)	(2.52)		
Firm Covered (treated) × Number of profiles (adoption) × Discretionary Accruals					-0.02905	-0.03010
					(-1.19)	(-1.23)
Firm F.E.	Yes	Yes	Yes	Yes	Yes	Yes
Industry-Year F.E.	Yes	Yes	Yes	Yes	Yes	Yes
Year-Month F.E.	Yes	Yes	Yes	Yes	Yes	Yes
Performance Controls	No	Yes	No	Yes	No	Yes
Cluster	Industry	Industry	Industry	Industry	Industry	Industry
Obs.	74,752	74,752	74,752	74,752	72,777	72,777
Adj. R-Squared	0.743	0.744	0.743	0.744	0.753	0.753

Sample period: 2009-2018.

The table reports the effects on shorting activity of different measures of manipulation and misreporting (REM and AEM). The period of interest is between the end of the fiscal year and the announcement of the results. The treatment group is defined by firms covered by the alternative data source, while the control group consists of the most similar firms based on textual similarity in filings (TNIC) that are not covered. The t-statistics are reported in parentheses (two-sided), standard errors are clustered at the industry level. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively. The industry classification is defined at the two-digit SIC code. Number of profiles is the total count of employees in securities trading firms (sophisticated investors) that have job descriptions compatible with "big-data". The outcome variable of interest is the short-interest index (SII). *RM* denotes the measure of real earnings management based on abnormally low discretionary expenses and cash flows from operations. *RM overproduction* is the measure of real earnings management based on overproduction. The controls include the market-to-book ratio, size, ROA, and NI scaled. Higher levels of real activity manipulation, but not accrual-based manipulation, are associated with increased shorting during the weeks between the fiscal year end and the announcement of the results for firms covered by the proxy when sophisticated investors have a high investment in human capital related to alternative data. The results also hold when controlling for the realized performance. The results are similar (although statistically weaker) to those presented in Table 2, indicating that the results discussed in the main analysis of this paper are not driven by the choice of interacting the measure of manipulation with the controls and fixed-effects.

.

Table A4. The Effects of Alternative Data on Earnings Response Coefficients

	(1) CAR	(2) CAR	(3) CAR sq	(4) CAR sq
Firm Covered (treated)	0.00500	0.00084	0.00719	0.00585
	(0.62)	(0.07)	(1.65)	(0.93)
Firm Covered (treated) × Number of profiles (adoption)	-0.00409**	-0.00322	-0.00269***	-0.00224*
	(-2.42)	(-1.44)	(-2.87)	(-1.80)
Firm Covered (treated) × Number of profiles (adoption) × SUE Score	0.00039		0.00028	
	(0.81)		(1.18)	
Firm Covered (treated) × Neg. Surprise Ind. × Number of profiles (adoption)		-0.00525		-0.00175
		(-1.38)		(-0.88)
Industry-Year F.E.	Yes	Yes	Yes	Yes
Year-month F.E.	Yes	Yes	Yes	Yes
Cluster	Industry	Industry	Industry	Industry
Obs.	9500	10592	9500	10592
Adj. R-Square	0.221	0.194	0.098	0.077

Sample period: 2009-2018.

The table reports the effects of data adoption on Earnings Response Coefficients. The period of interest is the three-day event window around the announcement of the results. The treatment group is defined by firms covered by the alternative data source, while the control group consists of the most similar firms based on textual similarity in filings (TNIC) that are not covered. The estimation follows the approach described in Gipper et al. (2019). The t-statistics are reported in parentheses (two-sided), standard errors are clustered at the industry level. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively. The industry classification is defined at the two-digit SIC code. Number of profiles is the total count of employees in securities trading firms (sophisticated investors) that have job descriptions compatible with "big-data". The outcome variable of interest is the cumulative abnormal returns (CAR), measured both in absolute terms and squared form. *SUE Score* (from IBES) is the ratio of the absolute surprise to estimate dispersion, and *Neg. Surprise Ind.* is a dummy variable taking the value of 1 if the firm announced negative unexpected earnings. Negative earnings surprises are associated with decreased market reactions for treated firms when investors are using the data, although the result is not statistically significant.

.

Table A5. The Effects of Alternative Data on REM & Accruals (no interaction)

	(1) Total Accruals	(2) RM Prod	(3) Delta RD	(4) CFO Scaled
Firm Covered (treated)	-0.59389**	-0.02569	-0.03261***	-0.01860
	(-2.49)	(-0.64)	(-3.52)	(-0.62)
Firm Covered (treated) × Number of profiles	0.07473	-0.00787**	0.00838**	0.01132**
	(1.38)	(-2.08)	(2.40)	(2.34)
Industry-Year F.E.	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes
Cluster	Industry	Industry	Industry	Industry
Obs.	4,677	9,277	7,448	9,530
Adj. R-Squared	0.890	0.694	0.066	0.456

Sample period: 2009-2018.

The table reports the main analysis of the paper: the relation between alternative data coverage and manipulation. T-statistics in parentheses (two-sided), standard errors are clustered at the industry level. The treatment group is defined by the coverage of the proxy used. The control is defined by the most similar firms based on textual similarity of product description in filings (TNIC) to those covered. the t-statistics are reported in parentheses (two-sided), standard errors are clustered at the industry level. *, ** , and *** indicate significance at the 10%, 5%, and 1% levels, respectively. Industry is defined at the two-digit SIC code. Number of profiles is the total count of employees in securities trading firms (sophisticated investors) that have job description compatible with "big-data". The coefficient of interest is the interaction of coverage and adoption (number of profiles). All estimations are run in a one-stage approach, with the relevant predictors of "normal" level (see Roychowdhury (2006)) included. Additional controls included in the estimation are: market-to-book ratio, size, institutional ownership, *ROA*, and leverage. The predicted signs, denoting the reduction of real activity manipulation stated in hypothesis 2, are: < 0 for *RM prod*, > 0 for Δ *R&D*, and > 0 for *CFO scaled*. The results indicate an overall decrease in real activity manipulation, with evidence of substitution towards AEM. The table differs from 5 as the results are estimated without the interaction of *number of profiles* with the control variables and the fixed effects.

Table A6. The Effects of Real Activity Manipulation on Post-Announcement Returns

	(1) Monthly Returns	(2) Monthly Returns	(3) Monthly Returns
Firm Covered (treated) × Number of profiles (adoption)	-0.00421	-0.00535	-0.00485
	(-1.36)	(-1.79)	(-1.00)
Firm Covered (treated) × Number of profiles (adoption) × RM	-0.00460**		
	(-2.64)		
Firm Covered (treated) × Number of profiles (adoption) × RM overproduction		-0.00357**	
		(-2.75)	
Firm Covered (treated) × Number of profiles (adoption) × Discretionary Accruals			0.00162
			(1.66)
Industry-Year F.E.	Yes	Yes	Yes
Year-Month F.E.	Yes	Yes	No
Performance Controls	Yes	Yes	Yes
Cluster	Industry	Industry	Industry
Obs.	48056	48056	46793
Adj. R-Squared	0.335	0.335	0.111

Sample period: 2009-2018.

The table reports the effects on stock returns of different measures of manipulation and misreporting (REM and AEM). The period of interest is the three months following the announcement of the results. The treatment group is defined by firms covered by the alternative data source, while the control group consists of the most similar firms based on textual similarity in filings (TNIC) that are not covered. the t-statistics are reported in parentheses (two-sided), standard errors are clustered at the industry level. *, ** , and *** indicate significance at the 10%, 5%, and 1% levels, respectively. The Industry classification is defined at the two-digit SIC code. Number of profiles is the total count of employees in securities trading firms (sophisticated investors) that have job description compatible with “big-data". *RM* denotes the measure of real earnings management based on abnormally low discretionary expenses and cash-flows from operations. *RM overproduction* indicates the measure of real earnings management based on overproduction. The controls include: market-to-book ratio, size, ROA, and NI scaled. Higher levels of real activity manipulation, but not accrual-based manipulation, are associated with lower returns for firms covered by the proxy when sophisticated investors employ higher levels of human capital related to alternative data. Appendix B (available upon request) reports similar results using the growth rate of profiles as a measure of adoption.

Table A7. The Effects of Alternative Data on AAERs and Restatements (Logit Estimation)

	(1) AAER Year	(2) AAER Year	(3) Restatement Year	(4) Restatement Year	(5) Big-R Year	(6) Big-R Year
Firm Covered (treated)	-5.79895**	-5.55964**	0.72301	0.55513	0.71041	0.40570
	(-2.29)	(-2.16)	(0.75)	(0.58)	(0.52)	(0.30)
Firm Covered (treated) × Number of profiles	0.97718**	0.90273*	-0.09614	-0.06434	-0.10653	-0.03761
	(2.11)	(1.88)	(-0.51)	(-0.35)	(-0.29)	(-0.11)
Model	Logit	Logit	Logit	Logit	Logit	Logit
Industry-Year F.E.	Yes	Yes	Yes	Yes	Yes	Yes
Controls	No	Yes	Yes	Yes	Yes	Yes
Cluster	Industry	Industry	Industry	Industry	Industry	Industry
Obs.	4944	3416	9136	6611	7346	5206
Pseudo R-Square	0.075	0.100	0.042	0.048	0.045	0.072

Sample period: 2009-2018.

The table reports the effects of alternative data coverage (treatment) on the likelihood of a firm facing an enforcement action by the SEC or a restatement. The t-statistics are reported in parentheses (two-sided), standard errors are clustered at the industry level. *, ** , and *** indicate significance at the 10%, 5%, and 1% levels, respectively. The treatment group is defined by the coverage of the proxy. The control is defined by the most similar firms in terms of TNIC similarity available in Compustat. Industry classification is defined at the two-digit SIC code. Number of profiles is the total count of employees in securities trading firms that have job descriptions compatible with “big-data", used as a proxy for investment by sophisticated investors in alternative data. The outcome variable is a dichotomous variable taking the value of 1 if a firm is the subject of an enforcement action by the SEC (AAER related to: revenues, inventory, accounts receivable, and COGS) or a restatement of interest in a given year. The results indicate that firms covered by alternative data have a higher likelihood of SEC enforcement actions. The findings are interpreted as additional evidence of the potential substitution from REM to AEM.

Table A8. The Effects of Alternative Data on AAERs and Restatements (robustness)

	(1) AAER Year	(2) AAER Year	(3) Restatement Year	(4) Restatement Year	(5) Big-R Year	(6) Big-R Year
Firm Covered (treated)	-0.00600**	-0.00769*	0.02026	0.02100	0.00668	0.00706
	(-2.28)	(-1.87)	(1.40)	(1.04)	(1.02)	(0.75)
Firm Covered (treated) × Number of profiles	0.00101*	0.00114	-0.00064	-0.00062	-0.00078	-0.00075
	(1.93)	(1.50)	(-0.25)	(-0.18)	(-0.54)	(-0.40)
Model	LinProbMod	LinProbMod	LinProbMod	LinProbMod	LinProbMod	LinProbMod
Industry-Year F.E.	Yes	Yes	Yes	Yes	Yes	Yes
Controls	No	Yes	Yes	Yes	Yes	Yes
Cluster	Industry	Industry	Industry	Industry	Industry	Industry
Obs.	53617	28877	53617	28877	53617	28877
R-Square	0.007	0.010	0.019	0.026	0.009	0.014

Sample period: 2009-2018.

The table reports the effects of alternative data coverage (treatment) on the likelihood of a firm facing an enforcement action by the SEC or a restatement. The t-statistics are reported in parentheses (two-sided), standard errors are clustered at the industry level. *, ** , and *** indicate significance at the 10%, 5%, and 1% levels, respectively. The treatment group is defined by the coverage of the proxy. The control is defined by the rest of non-utilities and non-financial firms available in Compustat. Industry classification is defined at the two-digit SIC code. Number of profiles is the total count of employees in securities trading firms that have job descriptions compatible with "big-data", used as a proxy for investment by sophisticated investors in alternative data. The outcome variable is a dichotomous variable taking the value of 1 if a firm is the subject of an enforcement action by the SEC (AAER related to: revenues, inventory, accounts receivable, and COGS) or a restatement of interest in a given year. The results indicate that firms covered by alternative data have a higher likelihood of SEC enforcement actions. The findings are interpreted as additional evidence of the potential substitution from REM to AEM.

Table A9. SEC investigations & Alternative Data

	(1) Active Case Indicator	(2) Active Case Indicator	(3) Active Case Indicator
Firm Covered (treated)	-0.10364	-0.06020	-0.47210
	(-0.19)	(-0.12)	(-0.71)
Firm Covered (treated) × Number of profiles (adoption)	0.14808	0.13603	0.19977
	(1.30)	(1.16)	(1.22)
Model	Logit	Logit	Logit
Industry-Year F.E.	Yes	Yes	Yes
Year-month F.E.	No	Yes	Yes
Controls	No	No	Yes
Cluster	Industry	Industry	Industry
Obs.	18059	17687	13420
Pseudo R-Square	0.068	0.074	0.128

Sample period: 2009-2017.

The table reports the results of a linear probability model regression where the outcome of interest is an indicator taking the value of 1 if the firm is under SEC investigation in a given month. I obtain raw data on all closed SEC investigations between January 1, 2000 and August 2, 2017 from Blackburne et al. (2021). The t-statistics are reported in parentheses (two-sided), standard errors are clustered at the industry level. *, ** , and *** indicate significance at the 10%, 5%, and 1% levels, respectively. The treatment group is defined by the coverage of the proxy used. The control is defined by the most similar firms based on textual similarity of product description in filings (TNIC) to those covered. Industry is defined at the two-digit SIC code. Number of profiles is the total count of employees in securities trading firms (sophisticated investors) that have job description compatible with “big-data". The coefficient of interest is the interaction of coverage and adoption (number of profiles). The results indicate an increase (although not statistically signficant) in the number of investigation by the SEC of firms treated when adoption is higher.

Table A10. Effects of Manipulation on Shorting Positions, Controlling for EPS surprise

	(1) SII	(2) SII	(3) SII	(4) SII	(5) SII	(6) SII
Firm Covered (treated)	-0.16720	-0.21522	0.44835	0.44696	-1.14853	-1.08329
	(2.24028)	(2.25848)	(2.38236)	(2.39204)	(2.61845)	(2.72373)
Firm Covered (treated) × Number of profiles (adoption)	0.31257	0.31781	0.17505	0.17329	0.57760	0.56358
	(0.46375)	(0.46676)	(0.49094)	(0.49143)	(0.54069)	(0.56023)
Firm Covered (treated) × Number of profiles (adoption) × RM	0.04917	0.06251*				
	(0.02806)	(0.03271)				
Firm Covered (treated) × Number of profiles (adoption) × RM overproduction			0.05137*	0.06411**		
			(0.02462)	(0.02665)		
Firm Covered (treated) × Number of profiles (adoption) × Discretionary Accruals					-0.20987	-0.18838
					(0.16407)	(0.17729)
Industry-Year F.E.	Yes	Yes	Yes	Yes	Yes	Yes
Year-Month F.E.	Yes	Yes	Yes	Yes	Yes	Yes
Additional Controls	NegSurpriseInd	SueScore	NegSurpriseInd	SueScore	NegSurpriseInd	SueScore
Cluster	Industry	Industry	Industry	Industry	Industry	Industry
Obs.	56415	56415	56415	56415	54548	54548
Adj. R-Squared	0.220	0.221	0.219	0.219	0.204	0.202

Sample period: 2009-2018.

The table reports the effects on shorting activity of different measures of manipulation and misreporting (REM and AEM). The period of interest is the one between the end of the fiscal year and the announcement of the results. Standard errors, clustered at the industry level, are provided in parentheses. *, ** , and *** indicate significance at the 10%, 5%, and 1% levels, respectively. The treatment group is defined by firms covered by the alternative data source, while the control group consists of the most similar firms based on textual similarity in filings (TNIC) that are not covered. The Industry classification is defined at the two-digit SIC code. Number of profiles is the total count of employees in securities trading firms (sophisticated investors) that have job description compatible with "big-data". The outcome variable of interest is the short-interest index (SII). *RM* denotes the measure of real earnings management based on abnormally low discretionary expenses and cash-flows from operations. *RM overproduction* is the measure of real earnings management based on overproduction. The controls include: market-to-book ratio, size, ROA, and NI scaled. Additionally, odd numbered columns also include a negative surprise indicator (taking the value of 1 if $suescore < 0$) control, while even numbered columns include suescore as a control. The results are qualitatively similar to those of the main analysis (Table 2), although statistically weaker, indicating that the main results are not entirely driven by the realized performance but by the monitoring of real activity manipulation. Table A13 reports a similar analysis, where the sample contains only the firms that have exhibited negative surprise.

.

Placebo Tests:

Table A11. Adoption (placebo) treatment: Effects of Manipulation on Shorting Positions

	(1) SII	(2) SII	(3) SII	(4) SII	(5) SII	(6) SII
Firm Covered (treat_placebo) × Number of profiles (adoption)	0.52223	0.60015	1.38771	1.30806	0.72887	0.61879
	(0.49)	(0.58)	(1.48)	(1.52)	(0.72)	(0.64)
Firm Covered (treat_placebo) × Number of profiles (adoption) × RM	-9.14495	-7.29580				
	(-1.47)	(-1.21)				
Firm Covered (treat_placebo) × Number of profiles (adoption) × RM overproduction			-2.63998	-1.67092		
			(-0.75)	(-0.49)		
Firm Covered (treat_placebo) × Number of profiles (adoption) × Discretionary Accruals					1.60610	1.67345
					(1.26)	(1.34)
Industry-Year F.E.	Yes	Yes	Yes	Yes	Yes	Yes
Year-Month F.E.	Yes	Yes	Yes	Yes	No	No
Performance Controls	No	Yes	No	Yes	Yes	Yes
Cluster	Industry	Industry	Industry	Industry	Industry	Industry
Obs.	30,343	30343	30,343	30,343	30,198	30,198
Adj. R-Squared	0.390	0.400	0.400	0.413	0.389	0.403

Sample period: 2009-2018.

The table presents results on the effects of different measures of manipulation and misreporting (REM and AEM) on shorting activity, when the treatment assignment is based on a placebo. The period of interest is between the end of the fiscal year and the earnings announcement. Standard errors are clustered at the industry level and reported in parentheses. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively. The industry classification is defined at the two-digit SIC code. The (placebo) treatment group is defined by a random sample of firms, excluding the coverage of the proxy of interest. The control group consists of the remaining non-treated (non-utility and non-financial) firms. The number of profiles refers to the total count of employees in securities trading firms that have job descriptions compatible with "big-data", serving as a proxy for sophisticated investor investment in alternative data. The outcome variable of interest is the short-interest index (SII). *RM* denotes the measure of real earnings management based on abnormally low discretionary expenses and cash flows from operations. *RM overproduction* is the measure of real earnings management based on overproduction. The controls include the market-to-book ratio, firm size, ROA, and scaled net income. Following Gipper et al. (2019), the measure of manipulation is interacted with both the controls and the fixed effects.

The table reports the results of a placebo test version of the tests discussed in Table 2. According to Hypothesis 1, which states that treated firms will face stronger monitoring, the predicted sign of the coefficient of interest (the triple interaction in columns (1) to (4)) is: $\beta > 0$.

The table shows that higher levels of real activity manipulation (columns (1) and (3)) are not associated with increased shorting during the weeks between fiscal year end and the earnings announcement when using a placebo treatment. The results are similar when controlling for realized performance (columns (2) and (4)).

.

Table A12. Adoption (placebo) treatment: effects on REM & Accruals

	(1) Total Accruals	(2) RM Prod	(3) Delta RD	(4) CFO Scaled
Firm Covered (treat_placebo)	-0.40437	-0.07924	0.11630	-0.04564
	(-0.18)	(-1.08)	(0.17)	(-0.64)
Number of profiles	7.25254	0.19009**	-0.12445	0.06389
	(1.45)	(2.57)	(-0.45)	(1.17)
Firm Covered (treat_placebo) × Number of profiles	-0.11757	0.01582	0.02297	0.01819
	(-0.28)	(0.94)	(0.14)	(1.50)
Industry-Year F.E.	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes
Cluster	Industry	Industry	Industry	Industry
Obs.	5,260	12,074	9,433	13,137
Adj. R-Squared	0.011	0.496	0.009	0.537

Sample period: 2009-2018.

The table reports the placebo treatment version of the main analysis, examining the relationship between alternative data coverage and manipulation when treatment assignment is randomized (placebo). The t-statistics are reported in parentheses (two-sided). Standard errors are clustered at the industry level. The (placebo) treatment group is defined by a random sample of firms, excluding the coverage of the proxy of interest. The control group consists of the remaining non-treated companies (excluding utilities and financials). Industry is defined at the two-digit SIC code level. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively. Number of profiles is the total count of employees in securities trading firms (sophisticated investors) whose job descriptions are compatible with "big data". The coefficient of interest is the interaction of coverage and adoption (number of profiles). All estimations use a one-stage approach including predictors of the normal level (see Roychowdhury (2006)). Additional controls included in the estimation are: market-to-book ratio, size, institutional ownership, *ROA*, and leverage. The predicted signs, reflecting the hypothesized reduction in real activity manipulation (Hypothesis 2), are: negative (< 0) for *RM prod*, positive (> 0) for Δ *R&D*, and positive (> 0) for *CFO scaled*. The results do not indicate a significant change in real activity manipulation or aggressive accounting manipulation.

Table A13. Effects of Manipulation on Shorting Positions (negative surprise)

	(1)	(2)	(3)	(4)	(5)	(6)
	SII	SII	SII	SII	SII	SII
Firm Covered (treated)	-9.92968	-9.60531	-8.49870	-8.44975	-1.97579	-1.53695
	(10.40047)	(9.89573)	(7.64034)	(7.85550)	(2.75330)	(2.75348)
Firm Covered (treated) × Number of profiles (adoption)	1.77650	1.83095	1.62561	1.68677	0.63929	0.47894
	(2.26706)	(2.24682)	(1.87164)	(1.91790)	(0.98505)	(1.00609)
Firm Covered (treated) × Number of profiles (adoption) × RM	1.94939	2.14820				
	(1.44030)	(1.43642)				
Firm Covered (treated) × Number of profiles (adoption) × RM overproduction			3.51726***	4.04273***		
			(0.64633)	(0.97376)		
Firm Covered (treated) × Number of profiles (adoption) × Discretionary Accruals					1.70830	2.49188
					(2.49403)	(2.04805)
Industry-Year F.E.	Yes	Yes	Yes	Yes	Yes	Yes
Year-Month F.E.	Yes	Yes	Yes	Yes	No	No
Performance Controls	No	Yes	No	Yes	Yes	Yes
Cluster	Industry	Industry	Industry	Industry	Industry	Industry
Obs.	11672	11672	11672	11672	11604	11604
Adj. R-Squared	0.683	0.694	0.688	0.701	0.559	0.565

Sample period: 2009-2018.

The table reports the effects on shorting activity of different measures of manipulation and misreporting (REM and AEM), when firms announcement led to negative surprise. The period of interest is the one between the end of the fiscal year and the announcement of the results. Standard errors, clustered at the industry level, are provided in parentheses. *, ** , and *** indicate significance at the 10%, 5%, and 1% levels, respectively. The treatment group is defined by firms covered by the alternative data source, while the control group consists of the most similar firms based on textual similarity in filings (TNIC) that are not covered. The Industry classification is defined at the two-digit SIC code. Number of profiles is the total count of employees in securities trading firms (sophisticated investors) that have job description compatible with “big-data". The outcome variable of interest is the short-interest index (SII). *RM* denotes the measure of real earnings management based on abnormally low discretionary expenses and cash-flows from operations. *RM overproduction* is the measure of real earnings management based on overproduction. The controls include: market-to-book ratio, size, ROA, and NI scaled. The sample is constructed by considering only observations with negative EPS surprise. Compared to the results presented and discussed in Table 2, the size of the coefficients is noticeably larger although the statistical significance is strongly reduced in column (1) and (2). The direction of the results has not changed, leading to a similar interpretation: even when conditioning the sample on negative earnings surprises, higher level of real activity manipulation lead to stronger shorting for treated firms.

.

Advertisement Appendix:
Summary Statistics: Advertisement Expenditures

Table A14. Summary Statistics: Monthly Advertisement Expenditures

	25th percentile	Median	75th percentile	Mean	Standard Deviation
Monthly Ads Expenditures	4571	77495.5	1432501	3705198	9279961
Observations	24256				

The table reports the raw measure of monthly expenditures in advertisement by the firms covered in the available sample.

.

Summary Statistics: Firms Covered vs Not

Market-to-Book ratio

	25th percentile	Median	75th percentile	Mean	Standard Deviation
Non-covered Firm	.6270092	1.610609	3.400504	2.185424	10.82274
Covered Firm (ads)	1.613191	2.959862	5.351389	4.307047	8.597732
Observations	48841				

Size

	25th percentile	Median	75th percentile	Mean	Standard Deviation
Non-covered Firm	3.021315	4.955318	7.003875	4.999114	2.719404
Covered Firm (ads)	5.796394	7.266633	8.677995	7.128856	2.181297
Observations	48995				

The table reports the summary statistics of MTB ratio and size for firms covered in the advertisement sample vs the rest of firms from Compustat used in the previous analyses.

.

	(1) SII	(2) SII	(3) Utilisation	(4) Utilisation	(5) Volume	(6) Volume
Month Ads Expenditure	-0.04445***	-0.02227***	-0.06123***	-0.06123***	-0.00897***	-0.00484*
	(-6.93)	(-2.84)	(-3.35)	(-3.35)	(-4.28)	(-1.76)
Lagged Month Ads Expenditure		-0.03075***	-0.03658**	-0.03658**		-0.00648**
		(-3.77)	(-2.07)	(-2.07)		(-2.20)
Industry-Month F.E.	Yes	Yes	Yes	Yes	Yes	Yes
Year-Month F.E.	Yes	Yes	Yes	Yes	Yes	Yes
Firm F.E.	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Cluster	IndustryMonth	IndustryMonth	IndustryMonth	IndustryMonth	IndustryMonth	IndustryMonth
Obs.	12,500	11,626	12,137	12,137	12,552	11,675
Adj. R-Squared	0.590	0.597	0.678	0.678	0.514	0.511

Advertisement Expenditures Analysis:

Table A15. Monthly Advertisement Expenditures and Stock Market reaction

Sample period: 2010-2019.

The table reports the results of the analysis on the effects of changes in advertisement expenditure patterns on capital markets. Month Ads Expenditure measures is the monthly advertisement expenditure as recorded in Ad-Intel measured at the firm level. The t-statistics are reported in parentheses (two-sided), standard errors are clustered at the industry-month level. *, ** , and *** indicate significance at the 10%, 5%, and 1% levels, respectively. Industry is defined using the two-digit SIC code. The coefficient of interest is the one associated with Month Ads Expenditure.

Higher levels of advertisement are associated with contemporaneous lower shorting activity, as measured by SII and Utilization. The results indicate that an increase in advertisement leads to reduced shorting of the stock (columns (1)-(4)), accompanied by a decrease in volume (columns (5) to (6)). The results are interpreted as follows: changes in advertisement patterns recorded through a granular source of advertisement expenditures (an alternative data source) are reflected in stock market reactions, thus lending credibility to the monitoring channel investigated in this paper.

Table A16. Abnormal Advertisement Expenditures and Stock Market reaction (long-version)

	(1) SII	(2) SII	(3) SII	(4) Utilisation	(5) Utilisation	(6) Utilisation	(7) Volume	(8) Volume	(9) Volume	(10) B/A Spread	(11) B/A Spread	(12) B/A Spread
Abnormal Ads Expenditure	-0.030***	-0.044***	-0.022***	0.006	-0.079***	-0.066***	-0.002	-0.009***	-0.005*	0.000**	0.000**	0.000*
	(-5.98)	(-6.93)	(-2.82)	(0.40)	(-5.34)	(-3.56)	(-1.15)	(-4.28)	(-1.88)	(2.14)	(2.37)	(1.77)
Lagged Abnormal Ads Expenditure			-0.031***			-0.030*			-0.006**			-0.000
			(-3.80)			(-1.69)			(-2.07)			(-0.39)
Industry-Month F.E.	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year-Month F.E.	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm F.E.	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes
Cluster	IndustryMonth	IndustryMonth	IndustryMonth	IndustryMonth	IndustryMonth	IndustryMonth	IndustryMonth	IndustryMonth	IndustryMonth	IndustryMonth	IndustryMonth	IndustryMonth
Obs.	18,846	12,500	11,626	19,761	13,059	12,137	18,899	12,552	11,675	21,961	14,988	13,943
Adj. R-Squared	0.706	0.590	0.597	0.607	0.669	0.678	0.464	0.514	0.511	0.497	0.484	0.551

Sample period: 2010-2019.

The table reports the results of the analysis on the effects of changes in advertisement expenditure patterns on capital markets. Abnormal Ads Expenditure measures deviation from normal behavior computed at the industry-month level based on expected monthly advertisement expenditure. Higher abnormal advertisement indicates ad spending above what is predicted for industry peers. The t-statistics are reported in parentheses (two-sided), standard errors are clustered at the industry-month level. *, ** , and *** indicate significance at the 10%, 5%, and 1% levels, respectively. Industry is defined using the two-digit SIC code. The coefficient of interest is the one associated with Abnormal Ads Expenditure.

Higher levels of advertisement are associated with contemporaneous lower shorting activity, as measured by SII and Utilization. The results indicate that an increase in advertisement leads to reduced shorting of the stock (columns (1)-(6)), accompanied by a decrease in volume and an increase in bid-ask spread (columns (7) to (12)). The results are interpreted as follows: deviations from expected behavior recorded through a granular source of advertisement expenditures (an alternative data source) are reflected in stock market reactions, thus lending credibility to the monitoring channel investigated in this paper.

Table A17. Abnormal Advertisement Expenditures, Stock Market reactions and Cost Shielding (alternative specifications)

	(1) SII	(2) SII	(3) Utilization	(4) Volume
No Cost Shielding × Abnormal Ads Expenditure	-0.020***	-0.040***	-0.068***	-0.004*
	(-3.38)	(-5.79)	(-3.70)	(-1.95)
Cost Shielding Indicator × Abnormal Ads Expenditure	-0.006	-0.022***	-0.054***	-0.001
	(-1.11)	(-3.86)	(-3.54)	(-0.60)
Industry-Month F.E.	Yes	Yes	Yes	Yes
Year-Month F.E.	Yes	Yes	Yes	Yes
Firm F.E.	Yes	Yes	Yes	Yes
Controls	No	Yes	Yes	Yes
Cluster	IndustryMonth	IndustryMonth	IndustryMonth	IndustryMonth
Obs.	9630	6352	6471	6358
Adj. R-Squared	0.746	0.621	0.679	0.623
T-Test Coeff	2.53	3.35	.96	2.07

Sample period: 2010-2019.

The table reports the results of the analysis of the effects of advertisement (as recorded in a source of alternative data) on capital markets. The t-statistics are reported in parentheses (two-sided), standard errors are clustered at the industry-month level. *, ** , and *** indicate significance at the 10%, 5%, and 1% levels, respectively. Abnormal Ads Expenditure is a measure of deviation from normal behavior computed at the industry-month level based on expected month advertisement expenditure. Industry is defined at the two-digit SIC code. *Cost Shielding Indicator* is a dummy variable taking the value of 1 if managers' compensation contracts include cost shielding provisions, defined as in Bloomfield et al. (2021). Higher level of advertisement are associated with contemporaneous lower shorting activity more for firms where the managers bear the cost of increasing advertisement. The key difference from Table 4 lies in how the "cost shielding" indicator variable enters into the regression. This table explicitly tests whether the financial market's reaction differs significantly between firms with and without cost shielding provisions, as evidenced by the statistically significant results reported in the T-test line.