Journal of Accounting and Economics 76 (2023) 101573

Contents lists available at ScienceDirect

Journal of Accounting and Economics

journal homepage: www.journals.elsevier.com/journal-of-accounting-and-economics

Lost in standardization: Effects of financial statement database discrepancies on inference[☆]

Kai Du [a], Steven Huddart [a, *], Xin Daniel Jiang [b]

[a] *Smeal College of Business, Penn State University, United States*
[b] *School of Accounting and Finance, University of Waterloo, Canada*

ARTICLE INFO

Article history:
Received 23 August 2021
Received in revised form 13 December 2022
Accepted 15 December 2022
Available online 26 December 2022

JEL classification:
C81
G14
M41

Keywords:
Accrual accounting
Real earnings management
Accounting-based anomalies
Disclosure quality
Data aggregators
XBRL

ABSTRACT

SEC-mandated, machine-readable structured filings are an alternative source to Compustat for companies' accounting data. Discrepancies between as-filed and Compustat data, potentially a result of Compustat's standardizations, are more pronounced for firms with complex financial reporting. We show that these data discrepancies affect inferences in four research settings: (i) properties of accrual accounting, including accruals-cash flow relationships and abnormal accruals; (ii) real earnings management; (iii) the existence and magnitude of six of 21 accounting-based anomalies examined, including the accruals anomaly; and (iv) disclosure quality assessments based on the hierarchical structure of financial statement items. FactSet data also exhibit significant and often larger discrepancies from as-filed data. Our findings demonstrate the importance of these data discrepancies for the interpretation of empirical tests.

© 2022 Elsevier B.V. All rights reserved.

1. Introduction

Financial statement data assembled by third-party, commercial data aggregators (e.g., Compustat from S&P Global) form the basis of numerous archival studies in accounting and finance as well as the trading decisions of investment professionals. Data aggregators' efforts to achieve consistency, over time and across companies, in the data they extract from SEC registrants'

[☆] We thank Bob Holthausen (editor), Joseph Engelberg (referee), Daniel Aobdia, Sam Bonsall, Changling Chen, James Choi, Lauren Cohen, Ying Compton, Michael Dambra, Chengfeng Du, Kurt Gee, Walter Hamscher, Byoung-Hyoun Hwang, Michael Kirschenheiter, Juhani Linnainmaa, Henock Louis, Julie Marlowe, Rick Mergenthaler, Brian Miller (FARS discussant), Karl Muller, Jed Neilson, Doron Nissim, Stephen Penman, Anup Srivastava, Inho Suk, Rebecca Tang, Ayung Tseng, John Turner, Shuyang Wang, Christine Wiedman, Mike Willis, Biqin Xie, Michael Zhang, Rong Zhao, and workshop participants at the Securities and Exchange Commission (DERA), Penn State, University at Buffalo, University of Calgary, University of Waterloo, XBRL US, the Accounting Design Project, and the 2022 FARS Midyear Meeting for helpful comments and suggestions. We are especially grateful to Dan Givoly for many discussions. We thank Shuping Chen and Bin Miao for sharing the disclosure quality data. We thank Alina Lerman and Frank Zhang for their help with the replication of their study. We acknowledge our respective institutions for financial support. The as-filed data used in this study are available through Wharton Research Data Services (WRDS).

[*] Corresponding author.
E-mail addresses: kxd30@psu.edu (K. Du), huddart@psu.edu (S. Huddart), daniel.jiang@uwaterloo.ca (X.D. Jiang).

https://doi.org/10.1016/j.jacceco.2022.101573
0165-4101/© 2022 Elsevier B.V. All rights reserved.

complex and voluminous filings have led aggregators to adopt standardization practices. Although intended to mitigate the impact of diverse financial reporting, these practices have raised concerns among corporate managers and regulators. Former Deputy Chief Accountant of the Securities and Exchange Commission (SEC) Julie Erhardt (2016) pointedly asks corporate financial managers, “do you know how the financial information provided by third parties compares with the financial statements that your company filed with the Commission?” Users of commercial data would benefit from knowing the nature and significance of discrepancies between financial statements as originally filed with the SEC and the data provided by aggregators.[1] We provide evidence that discrepancies between SEC filings and data aggregators' products are widespread and large enough to affect inferences and decision-making.

Until recently, it has been infeasible to systematically evaluate the implications of data aggregators' standardization practices because no comprehensive, machine-readable “as-filed” financial statement data existed. The situation changed with the advent of SEC-mandated structured disclosures in eXtensible Business Reporting Language (XBRL) format, designed to facilitate automated data retrieval and analysis (SEC, 2009).[2]

Structured disclosures are based on the U.S. Generally Accepted Accounting Principles (GAAP) Financial Reporting Taxonomy. Financial statement data extracted directly from XBRL filings have been heavily used by the SEC to support its oversight efforts, including fraud detection and risk monitoring (PwC, 2014; Merrill Corporation, 2016). Interest in using XBRL data is also growing in the investment community (Willis, 2013).

We use the newly available structured disclosures to evaluate the implications of data aggregators' standardization practices for archival research using financial statement data. We assemble as-filed financial statement data analogous to Compustat data items for 20,410 firm-years from 2012 to 2019. Our analysis focuses on four research settings: (i) properties of accounting accruals; (ii) real earnings management; (iii) the existence and magnitude of 21 accounting-based anomalies, including the accruals anomaly; and (iv) implications of disclosure quality (DQ) as measured using the hierarchical structure of financial statement items.

There are three advantages to as-filed data, as opposed to third-party data provided by data aggregators. First, as-filed data are more granular than aggregators' data.[3] Second, as-filed data adhere to an authoritative, public taxonomy. They are not subject to a data aggregator's proprietary adjustments and thus are verifiable and reproducible. Third, whereas aggregators may take days or even weeks to compile financial statement data and make them available to investors, as-filed data are available as soon as the XBRL filings are submitted to the SEC.[4]

Before examining specific research settings, we first document significant discrepancies between the as-filed and Compustat data for 68 out of 73 data items frequently used in archival research. These discrepancies tend to be greater when (i) the filing contains more granular items that are located deeper in the hierarchy defined by FASB's taxonomy; (ii) the filing contains more tags, a larger fraction of industry-specific tags, or a larger fraction of custom tags; or (iii) comparison of the accounting practices between industry peers is more difficult. In other words, when a firm has complex accounting or discloses uncommon accounting items, Compustat data tend to deviate more from as-filed data.[5]

Our first set of analyses re-examines three well-studied properties of accrual accounting. First, the timing role of accrual accounting implies that contemporaneous accruals and cash flows are negatively correlated (Dechow, 1994). Bushman et al. (2016) document that the negative contemporaneous correlation has attenuated over time, which they attribute to increases in non-timing-related accrual recognition. In the period for which as-filed data are available, the negative correlation is significantly stronger in as-filed data than in Compustat data. Therefore, the attenuation in the negative correlation documented in prior research may be partly due to data discrepancies instead of non-timing-related accruals (e.g., one-time items).

Second, we re-examine how accruals contribute to earnings' ability to predict future cash flows, on which prior research has offered mixed evidence (Bushman et al., 2016; Nallareddy et al., 2020). Nallareddy et al. (2020) show that the incremental predictive ability of cash flows (accruals) almost monotonically increases (decreases) over their sample period, which they attribute to changes in firms' operating environment. Indeed, we show that using Compustat data, the incremental predictive power of accruals has disappeared in recent years. However, using as-filed data, accruals retain incremental predictive power. In other words, data discrepancies account for part of the decline in the predictive ability of accruals.

Third, we revisit the measure of abnormal accruals based on estimating an accrual model in the cross-section of each industry (Dechow and Dichev, 2002). By adopting this measure as a proxy for earnings quality, hundreds of studies have

[1] In a white paper by S&P Global, the data vendor states that “different data providers use different methodologies for standardization, and those methodologies have a definite impact on the presented data” (S&P Global, 2018, p. 2).

[2] In 2009, the SEC adopted a final rule requiring public companies to provide XBRL versions of their quarterly and annual financial reports in addition to a standard text or html filing (SEC, 2009).

[3] Compustat's Fundamental Annual dataset contains about 900 data items. By comparison, according to the 2020 edition of the XBRL taxonomy, there are 643 unique balance sheet tags, 574 income statement tags, and 766 cash flow statement tags. These counts pertain only to the numerical portion of the financial statements. A typical XBRL filing also contains many disclosure tags that tie text passages to specific disclosure topics (e.g., inventory policies).

[4] D'Souza et al. (2010) report that the median dissemination lag by Compustat is 15 weekdays, with the inter-quartile value ranging from 8 to 23 weekdays.

[5] In the data on the as-filed structure of financial statements, which we make publicly available, we include *firm-year* statistics of the level of depth of the three financial statements, which are indicative of the accounting complexity of the firm. We also provide a *data-item* statistic of complexity—the level of depth of the tag corresponding to the data item in Table IA.4. These depth statistics are an indication of which firms or data items are more susceptible to data discrepancies.

examined its determinants and consequences (Dechow et al., 2010). We show that the explanatory power of the accrual model using as-filed data is 37% greater than using Compustat, and discrepancies in estimated abnormal accruals tend to be larger for smaller firms.

Our second set of analyses re-examines the finding that "suspect firms" tend to have greater real earnings management (Roychowdhury, 2006). We find that as-filed data yield significantly different coefficients on an indicator variable for suspect firms for two out of three measures of real earnings management: abnormal cash flows and abnormal discretionary expenses. Strikingly, for abnormal discretionary expenses, we obtain the opposite signs for the coefficient.

Our third set of analyses focuses on accounting-based stock return anomalies, which have been incorporated into investment strategies and scrutinized by hundreds of published studies (Green et al., 2017; Linnainmaa and Roberts, 2018; Hou et al., 2020). We start by examining whether Compustat and as-filed data yield the same inferences about the existence and magnitude of the accruals anomaly (Sloan, 1996). Based on both portfolio analysis and Fama-MacBeth cross-sectional regressions, we show that when using Compustat data, *no* accruals anomaly is detected during the period for which as-filed data are available; however, when using as-filed data, the accruals anomaly *is* significant (i.e., the low-accruals portfolio has significantly higher returns than the high-accruals portfolio).

We then revisit 20 other accounting-based anomalies examined by Green et al. (2017) and Hou et al. (2020). We find that the discrepancies affect the predictive power of five other accounting-based return predictors: abnormal operating accruals (Xie, 2001), earnings before depreciation and extraordinary items-to-debt ratio (Ou and Penman, 1989), growth in long-term net operating assets (Fairfield et al., 2003), operating profitability (Fama and French, 2015; Ball et al., 2016), and taxable income (Lev and Nissim, 2004). These five predictors, as well as operating accruals, relative to the remaining 15 predictors, tend to involve data items that are deeper in the reporting taxonomy.

After documenting that the discrepancies in financial statement *values* between Compustat and as-filed data are large enough to affect inference, we turn to the differences in the *structure* of as-filed data relative to Compustat in the fourth set of analyses. Chen et al. (2015) propose a measure of DQ that captures the level of disaggregation of accounting data through a count of non-missing Compustat line items. We develop an analogous DQ measure based on the structure of XBRL tags reported according to the FASB taxonomy, instead of Compustat's balancing model. We show that this measure is positively correlated with the Compustat-based DQ measure and performs better in at least one validation test.

Lastly, we explore whether our findings generalize to another data aggregator by using FactSet data in place of Compustat data in our analyses. FactSet's discrepancies tend to be larger and more frequent than Compustat's discrepancies. We also replicate the analysis in specific research settings using FactSet and find different inferences using FactSet versus using as-filed data.

Implications for archival researchers. For archival researchers relying on financial statement data from commercial aggregators, our findings provide a cautionary note that discrepancies between commercial data and originally-filed accounting numbers are both common and large. The discrepancies are greater when the financial statement is more complex, and in certain industries (e.g., energy, shops, and durables). Moreover, different aggregators (e.g., Compustat and FactSet) exhibit non-overlapping discrepancies with as-filed data, implying that their standardization practices diverge.

In light of these discrepancies, we recommend using as-filed data in research that relies on data items with large discrepancies, especially when pre-2011 data are inessential for the research question.[6] Because the time and effort involved in preparing the as-filed data for analysis may pose a barrier to adoption, we make the as-filed analog of the Compustat data publicly available through Wharton Research Data Services (WRDS). Along with the Compustat-like data, we also provide a more granular set of as-filed tags prepared using our methodology, which may facilitate research that requires data items to be more flexibly defined than those provided in conventional databases. Also made available are data on the structure of as-filed financial statements, including disclosure quality, number of tags, and level of depth.

More importantly, we identify several research settings in which as-filed data are particularly preferable or, at a minimum, should be used for robustness checks. First, various temporal trends (i.e., diminishing negative correlation between accruals and cash flows, declining ability of accruals to predict cash flows, and attenuation of the accruals anomaly), which have been interpreted as arising from changes in reporting and business environments or investment practice, are partly attributable to Compustat's standardization practices. Second, empirical constructs (e.g., abnormal discretionary expenses) that heavily rely on data items exhibiting large discrepancies between data sources should be revisited using as-filed data. Third, in research designs where inference is affected by how observations are ordered and sorted into extreme portfolios (e.g., hedge portfolio analysis), the choice of data source may have a significant impact.[7] Lastly, research that focuses on the hierarchical relationships among Compustat's data items may also benefit from using as-filed data.

Section 2 recaps related research. Section 3 explains how we assemble as-filed financial statement data and quantifies discrepancies in Compustat data. Sections 4–7 show how inference depends on the data source in four settings: properties of accrual accounting, real earnings management, accounting-based stock return anomalies, and disclosure quality. In Section 8, we replicate these analyses using FactSet. Section 9 concludes.

[6] We acknowledge that the different conclusions drawn in various research settings may be specific to the post-2012 period for which as-filed data are available.

[7] For example, about 32% of the stocks in a portfolio formed using as-filed operating accruals are distinct from stocks in the corresponding portfolio formed using Compustat. Differences in portfolio composition affect portfolio returns and consequent inferences.

2. Related literature

2.1. Research on data aggregation

Data aggregators parse regulatory filings and other public information to produce ready-made datasets for practitioners and academics.[8] Several studies document the associations between data aggregators' dissemination of financial information and market participants' reactions to such information (e.g., D'Souza et al., 2010; Schaub, 2018; Akbas et al., 2018; Bochkay et al., 2022). In addition, a growing literature examines the integrity or quality of data aggregation products. In the case of Compustat, prior studies have examined the implications of survivorship bias (Davis, 1994), considered the lack of private firm coverage for research on industry concentration (Ali et al., 2008), and investigated its difference from Value Line (Kern and Morris, 1994). Prior research has also scrutinized other data products, including analyst forecast data from I/B/E/S (Payne and Thomas, 2003; Kaplan et al., 2021) and mutual fund performance data from Morningstar (Chen et al., 2021).

Prior research on discrepancies between Compustat and financial statements filed with the SEC includes Chychyla and Kogan (2015) and Boritz and No (2020). Chychyla and Kogan (2015) examine whether 30 accounting line items differ between Compustat and 10-K filings between October 2011 and September 2012. For a random sample of 105 firms, Boritz and No (2020) compare financial statement items with embedded XBRL data and data from Compustat and two other aggregators. They report that 100% of the items on the face of the 10-K filings' financial statements are also present in the XBRL data and that the values agree 99.9% of the time, which establishes the high quality of XBRL data. Both studies report that (i) challenges exist when mapping financial statements to aggregators' data and (ii) discrepancies between financial statements and aggregators' data are frequent, statistically significant, and often material.

We advance this line of inquiry in three ways. First, we improve on prior mapping efforts by making extensive use of the FASB's calculation linkbase.[9] This linkbase organizes monetary elements by specifying values at one level of the taxonomy that may be aggregated (via addition and subtraction) to arrive at values at the next higher level. This aggregation process is important because the SEC filing manual directs filers to use the element with the narrowest definition when there is a choice among different elements that have definitions consistent with a set of facts in one or more periods. Moreover, filers are directed to use the most specific type attribute and the most specific reference to authoritative guidance, such as to a specific paragraph in the accounting standard.[10] Because the taxonomy is designed to facilitate aggregation, values for upper-level concepts not explicitly provided in the filing can be inferred by aggregating the relevant lower-level values. Different from prior research, our method traverses the structured disclosures and imputes values not explicitly provided in the filing according to the calculation linkbase. Our method additionally handles instances where the filer uses custom tags by taking advantage of the calculation linkbase in the extension taxonomy. Following our methodology, a researcher can assemble as-filed data on a large scale tailored to their needs.

Second, we create a comprehensive as-filed dataset that covers a longer and more recent sample period.[11] Using this data, we quantify discrepancies in 73 Compustat items, many of which have been used in hundreds of research studies. We identify characteristics of the filing and the data item associated with greater discrepancies. We show that discrepancies are not only frequent and statistically significant, but also large enough to affect inference in three prominent research settings. Moreover, this dataset is publicly available to enable researchers to check whether the inference is affected in their own empirical setting.

2.2. Literature related to specific research settings

Four strands of literature are related to the specific research settings in our paper, respectively: (i) research on properties of accrual accounting, including the contemporaneous relationship between accruals and cash flows (Bushman et al., 2016), cash flow predictability (Nallareddy et al., 2020), and measurement of abnormal accruals (Dechow et al., 2010); (ii) the real earnings management literature (Roychowdhury, 2006); (iii) studies of accounting-based anomalies, as synthesized by Richardson et al. (2010), Green et al. (2017), and Hou et al. (2020)[12]; and (iv) research that uses the structure of data items to measure DQ (Chen et al., 2015).

Whereas prior studies in these areas have invariably used Compustat data, we show that conclusions depend on the data source. For example, prior research documents temporal trends in accrual-cash flow relationships (Bushman et al., 2016; Nallareddy et al., 2020), greater real earnings management among suspect firms (Roychowdhury, 2006), and gradual attenuation of the accrual anomaly (Green et al., 2011). As-filed data indicate different or more nuanced fact patterns.[13]

[8] Section 5.1 of Blankespoor et al. (2020) reviews research on data aggregation.

[9] A taxonomy defines tags that identify a datum, its attributes, and its relationships to other data. The taxonomy is available at https://www.fasb.org/xbrl.

[10] Edgar Filing Manual – Volume II, Version 61, March 2022, §6.6.26 ff.

[11] Prior studies examine more limited sets of data items. Bostwick et al. (2016) focus on cost of goods sold; Chychyla and Kogan (2015) examine 30 financial statement line items; Boritz and No (2020) focus on data items from Compustat for which there is a single corresponding tag in 10-K XBRL filings.

[12] Recent research in this area examines whether anomalies attenuate as more capital is deployed in trading strategies designed to exploit the anomaly (Green et al., 2011; McLean and Pontiff, 2016).

[13] Data discrepancies explain a portion of the temporal shifts in empirical regularities that have been attributed to changes in non-timing-related accrual recognition (Bushman et al., 2016), operating environment (Nallareddy et al., 2020), and hedge fund exploitation (Green et al., 2011).

3. Data and descriptive statistics

3.1. As-filed financial statement data vs. Compustat

Our "as-filed" data are based on the Financial Statement and Notes Data Sets compiled by the SEC, which contain financial statement information directly extracted from periodic corporate XBRL filings.[14] For some firms, data are available from 2009; however, we focus on the period of 2012–2019, during which *all* public firms were required to submit periodic filings in XBRL format.[15] When more than one annual filing (10-K or 10-K/A) exists for the same fiscal year, we use the most recent filing for each fiscal year (for analyses other than anomalies), or the most recent filing before the portfolio formation date for each year (for anomalies analysis).[16]

A major step of our data preparation process involves constructing an as-filed data set comparable to the annual fundamental file compiled by Compustat. We focus on 73 Compustat data items that are either (i) frequently used in accounting and finance research or (ii) key elements on one of the three financial statements. These 73 data items include, but are not limited to, those used in later analyses of specific research settings. Our list covers most data items frequently used in prior research. In Table IA.20, we tabulate the data items included in our study with more than 100 Google Scholar citations.

For each data item, we identify the highest-level tags in the taxonomy that correspond to the Compustat data items based on Compustat's balancing model for financial statement items (S&P Global, 2018). The mapping is unambiguous and is not subject to the researcher's discretion. Nevertheless, we validate this mapping by verifying that the selected tag (or the combination of several tags) dominates all other tags when following the procedure detailed in Appendix C.1.

Occasionally, however, the filing does not report a value for a high-level tag. In such cases, we use the hierarchical relations specified by the calculation linkbase to impute the high-level tag value from the values of the appropriate child tags. We then validate imputed tag values using several basic accounting identities. Appendix C.1 contains further details on these steps. We emphasize that these steps do not involve subjective judgments on our part because the linkbase encodes relationships among all standard tags as determined by the taxonomy.

When the standard taxonomy does not accommodate unique circumstances in a filer's disclosure, filers are permitted to extend the taxonomy by using custom tags in addition to the standard tags prescribed by the taxonomy.[17] We use an algorithm detailed in Appendix C.2 to search for the nearest equivalent standard tag for every custom tag. In essence, the algorithm seeks the best match (through fuzzy matching of tag labels) for each custom tag among the standard tags that are descendants of the same parent as the custom tag.[18]

Table 1 presents the descriptive statistics of two versions of the financial statement data. Among the 73 data items, 43 come from the balance sheet (Panel A), 12 come from the income statement (Panel B), and 18 come from the cash flow statement (Panel C). Each data item is scaled by total assets, which are essentially identical between the two data sources. We define data discrepancy (*Discr*) as the absolute value of the difference between Compustat and as-filed values. We find a statistically significant *Discr* for 39 out of 43 balance sheet items, all 12 income statement items, and 17 out of 18 cash flow statement items. In total, 68 or 93.2% of all examined data items exhibit a significant discrepancy.

Turning to the signed difference between the two data sources (*Diff_Mean*), we find that for 23 data items, the as-filed version is significantly larger than Compustat, whereas for 32 data items, the Compustat version is significantly larger than the as-filed counterpart. We also report the *z*-statistics from the Wilcoxon rank-sum test that examines whether a data item is significantly different between the two versions. The results reject the null hypothesis that both versions of the data are the same for 30 out of 73 data items at the significance level of 0.05. Collectively, Table 1 suggests that data discrepancy is pervasive among popular Compustat data items in all three financial statements (i.e., balance sheet, income statement, and cash flow statement).

3.2. Potential factors associated with data discrepancies

The discrepancy between Compustat and as-filed financial statement data is likely due to Compustat's standardization process, which includes adjustments, aggregations, and omissions. In support of this conjecture, we make four observations. First, according to a white paper by S&P Global, Compustat makes numerous adjustments, some of which are intended to

[14] See https://www.sec.gov/dera/data/financial-statement-and-notes-data-set.html. Our analysis focuses on annual financial statements on Form 10-K. However, XBRL filings are also available for other form types (e.g., 10-Q, 8-K). All XBRL filings contain a reference to the taxonomy under which they are prepared. The allowable taxonomies for 10-K and 10-Q filings are the same, so our methodology will generalize to quarterly financial statements.

[15] The 2009 SEC rule prescribes implementation in three phases: large accelerated filers submit in XBRL format for fiscal periods ending on or after June 15, 2009; all other large accelerated filers submit in XBRL format for fiscal periods ending on or after June 15, 2010; and all remaining filers submit in XBRL format for fiscal periods ending on or after June 15, 2011 (SEC, 2009).

[16] We examine the potential impact of amended filings and restated financial statements in Section 6.3.1.

[17] The SEC has acknowledged that the use of unnecessary custom tags could potentially reduce the comparability of inter-company data and has specified the limited circumstances under which a filer may use custom tags. See, for example, https://www.sec.gov/structureddata/announcement/100721-use-custom-tags.

[18] Our results are qualitatively the same without incorporating custom tags.

Table 1
Descriptive statistics for discrepancies between Compustat and as-filed data.

Panel A: Balance sheet

Data item	*As-filed Data*		*Compustat Data*		*Diff_Mean*	*Discr*	*Wilcoxon z-stat.*
	Mean	Median	Mean	Median			
Assets:							
[1] AT	–	–	–	–			
[2] ACT	0.520	0.508	0.520	0.508	0.000	0.000***	−0.000
[3] CHE	0.243	0.138	0.244	0.139	−0.001***	0.000***	−0.201
[4] CH	0.195	0.115	0.195	0.115	−0.000	0.000***	0.000
[5] IVST	0.050	0.000	0.051	0.000	−0.001***	0.000***	−0.614
[6] RECT	0.112	0.089	0.117	0.095	−0.005***	0.005***	−4.673***
[7] INVT	0.082	0.025	0.083	0.026	−0.001***	0.000***	−0.271
[8] ACO	0.106	0.035	0.093	0.028	0.013***	0.013***	20.973***
[9] XPP	0.006	0.000	0.006	0.000	−0.000***	0.000***	−0.034
[10] ACOX	0.094	0.027	0.080	0.021	0.014***	0.013***	22.666***
[11] ANCT	0.480	0.492	0.480	0.492	−0.000	0.000***	0.000
[12] PPENT	0.205	0.127	0.194	0.112	0.011***	0.134***	6.318***
[13] PPEGT	0.422	0.264	0.412	0.276	0.010***	0.117***	−3.799***
[14] DPACT	0.216	0.127	0.217	0.128	−0.001***	0.001***	−0.832
[15] IVAEQ	0.003	0.000	0.003	0.000	0.000***	0.000***	2.098**
[16] IVAO	0.007	0.000	0.008	0.000	−0.001***	0.001***	−14.690***
[17] INTAN	0.178	0.077	0.179	0.079	−0.001***	0.002***	−0.523
[18] INTANO	0.074	0.023	0.076	0.025	−0.002***	0.002***	−1.400
[19] GDWL	0.104	0.020	0.104	0.020	0.000***	0.000***	0.000
[20] AO	0.123	0.047	0.095	0.022	0.028***	0.147***	5.744***
Liabilities and shareholders' equity:							
[21] LT	0.828	0.527	0.828	0.527	−0.000	0.000***	−0.000
[22] LCT	0.453	0.206	0.453	0.206	−0.000	0.000***	−0.000
[23] DLC	0.090	0.006	0.080	0.007	0.010***	0.012***	−0.340
[24] DD1	0.048	0.003	0.035	0.002	0.013***	0.015***	10.049***
[25] NP	0.007	0.000	0.007	0.000	−0.000***	0.001***	2.811**
[26] AP	0.102	0.050	0.103	0.051	−0.001***	0.001***	−1.702*
[27] TXP	0.002	0.000	0.002	0.000	0.000***	0.000***	6.354***
[28] LCO	0.291	0.106	0.271	0.098	0.020***	0.037***	8.450***
[29] XACC	0.070	0.042	0.079	0.051	−0.009***	0.018***	−13.838***
[30] LCOX	0.176	0.043	0.150	0.029	0.026***	0.046***	13.243***
[31] LNCT	0.348	0.225	0.348	0.225	0.000**	0.000***	0.006
[32] DLTT	0.191	0.099	0.201	0.116	−0.010***	0.010***	−3.726***
[33] TXDITC	0.003	0.000	0.003	0.000	0.000	0.000	0.000
[34] TXDB	0.003	0.000	0.003	0.000	0.000	0.000	0.000
[35] ITCB	0.000	0.000	0.000	0.000	0.000	0.000	0.000
[36] LO	0.166	0.057	0.131	0.029	0.035***	0.036***	35.142***
[37] TEQ	0.172	0.473	0.172	0.473	−0.000	0.000***	0.000
[38] SEQ	0.169	0.469	0.169	0.469	−0.000	0.000***	0.000
[39] CEQ	0.180	0.473	0.180	0.473	−0.000	0.000***	0.000
[40] CSTK	0.085	0.001	0.085	0.001	0.000	0.000***	−0.084
[41] RE	−4.210	−0.135	−4.221	−0.150	0.011***	0.022***	1.941*
[42] PSTK	0.002	0.000	0.002	0.000	−0.000***	0.000***	3.457***
[43] MIBN	0.003	0.000	0.003	0.000	0.000***	0.000***	0.000

Panel B: Income statement

Data Item	*As-filed Data*		*Compustat Data*		*Diff_Mean*	*Discr*	*Wilcoxon z-stat.*
	Mean	Median	Mean	Median			
[44] SALE	0.862	0.688	0.875	0.699	−0.013***	0.015***	−2.503**
[45] COGS	0.597	0.403	0.610	0.421	−0.013***	0.060***	−4.245***
[46] XSGA	0.535	0.272	0.551	0.277	−0.016***	0.057***	−2.399**
[47] XRD	0.105	0.007	0.105	0.007	−0.000***	0.000***	−0.033
[48] XAD	0.012	0.000	0.012	0.000	0.000***	0.000***	0.179
[49] DP	0.038	0.030	0.042	0.034	−0.004***	0.005***	−13.774***
[50] AM	0.008	0.001	0.008	0.002	−0.000***	0.000***	−1.377
[51] OIADP	−0.422	0.010	−0.400	0.026	−0.022***	0.026***	−7.799***
[52] XINT	0.056	0.008	0.058	0.010	−0.002***	0.003***	−8.330***
[53] PI	−0.457	−0.000	−0.453	0.000	−0.004***	0.009***	−0.007

Table 1 (*continued*)

Panel B: Income statement

Data Item	*As-filed Data*		*Compustat Data*		*Diff_Mean*	*Discr*	*Wilcoxon z-stat.*
	Mean	Median	Mean	Median			
[54] TXT	0.010	0.002	0.010	0.002	−0.000***	0.000***	−0.006
[55] IB	−0.605	−0.004	−0.605	−0.004	−0.000***	0.002***	−0.132

Panel C: Cash flow statement

Data item	*As-filed Data*		*Compustat Data*		*Diff_Mean*	*Discr*	*Wilcoxon z-stat.*
	Mean	Median	Mean	Median			
Operating activities:							
[56] OANCF	−0.159	0.051	−0.171	0.050	0.012***	0.021***	2.704**
[57] IBC	−1.010	−0.004	−1.010	−0.005	0.001	0.020***	0.885
[58] OPCAPCH	−0.036	0.003	−0.035	0.003	−0.001***	0.006***	−1.120
[59] RECCH	−0.006	−0.002	−0.007	−0.002	0.001***	0.003***	2.452**
[60] INVCH	−0.004	0.000	−0.004	0.000	0.000***	0.001***	1.607
[61] TXACH	0.000	0.000	0.000	0.000	0.000***	0.000***	0.189
[62] APALCH	0.041	0.004	0.044	0.005	−0.003***	0.005***	−3.227***
[63] AOLOCH	0.003	−0.000	0.003	−0.000	−0.000	0.014***	−1.476
[64] DPC	0.042	0.033	0.045	0.036	−0.003***	0.006***	−14.084***
[65] XIDOC	−0.000	0.000	−0.000	0.000	0.000	0.000	0.000
Investing activities:							
[66] IVNCF	−0.064	−0.041	−0.065	−0.041	0.001***	0.002***	−0.039
[67] CAPX	0.038	0.020	0.043	0.023	−0.005***	0.005***	−7.518***
[68] SPPE	0.002	0.000	0.002	0.000	−0.0001***	0.0003***	0.543
[69] AQC	0.020	0.000	0.022	0.000	−0.000	0.002***	0.667
Financing activities:							
[70] FINCF	0.330	0.006	0.329	0.006	0.001**	0.002***	0.075
[71] SSTK	0.163	0.004	0.175	0.004	−0.008***	0.040***	−1.552
[72] PRSTKC	0.016	0.000	0.016	0.000	−0.001***	0.001***	−2.017**
[73] DV	0.009	0.000	0.010	0.000	−0.0001***	0.001***	3.697***

Table 1 shows the data discrepancies between as-filed and Compustat values for 73 data items. See Table IA.4 for the definition of Compustat mnemonics. All variables are scaled by the total assets. *Diff_Mean* (*Discr*) is the mean of *signed* (*unsigned*) difference between the as-filed and Compustat values. We also report the *z*-statistics from the Wilcoxon rank-sum test. The sample consists of 20,410 firm-year observations over 2012−2019. Significance at the 10%, 5%, and 1% levels (two-sided) are denoted by *, **, and ***, respectively. Indented data items are components of higher-level data items.

remove variations in a same-firm datum over time (S&P Global, 2018).[19] Second, a comparison of the definitions of Compustat's data items with those of the corresponding XBRL tags reveals that Compustat makes several adjustments to reported values in financial statements. Third, private communications with the technical staff of S&P Global confirm that Compustat does not utilize XBRL filings. Instead, Compustat staff must read filings and, guided by a proprietary data collection manual, assemble data by selecting and combining disclosed values from the filings (see Appendix B). This process necessarily involves subjective judgment. Finally, we note that Compustat may misclassify or omit financial statement items during the standardization process, as illustrated by the two examples in the Internet Appendix (Tables IA.1 and IA.2). In the following, we consider several factors that may affect the magnitude of the data discrepancy.

Financial statement depth (*Depth*). Compustat aggregates financial statement line items to give a more standardized menu of data items. Regarding the extent to which this “bottom-up” aggregation involves discretion, we suspect that the magnitude of adjustments made by Compustat is correlated with the level of disaggregation of the original filing. A characteristic of each tagged value in as-filed financial statement data is its location in a hierarchy of parent-child relationships that culminate in a high-level aggregate value, such as total assets. We measure the level of disaggregation by *Depth*, the average level of “depth” in the hierarchy of all XBRL tags reported in the three financial statements. A greater average depth presents a greater challenge for the Compustat staff, who aggregate raw items into standardized data items.

Total number of tags (*NTag*). The total number of XBRL tags has been used as a proxy for accounting complexity (Hoitash and Hoitash, 2018). Compustat's standardization process may involve more discretion and errors for complex financial statements.

Industry-specific tags (*IndTag*). Compustat's standardization process may lead to larger discrepancies when the accounting standards applicable to a firm's business are highly industry-specific. We measure industry specificity by the percentage of industry-specific tags (*IndTag*) on the three financial statements. Industry-specific tags are tags that refer to Accounting Standards Codification Topic Area 900, which provides guidance for specific industries (e.g., airlines) or activities (e.g., mining).

[19] “There is a certain amount of ‘noise’ or variation in data that a highly standardized data source will try to avoid. … By removing more noise through standardization, Compustat delivers more consistent data. In general, Compustat's standardization practices lead to cleaner quantitative models that require less correction for outliers.” See pp. 10−12 of S&P Global (2018).

Custom tags (*CustomTag*). Custom tags are indicative of firm-specific reporting situations not covered by the standard FASB taxonomy. These situations lead to nonstandard financial statement items and create challenges for Compustat's standardization process. Therefore, we expect the discrepancy to be greater for firms with more custom tags. The prevalence of custom tags (*CustomTag*) is measured as the percentage of custom tags on the three financial statements.

Accounting comparability (*CompAcctInd*). The next variable we examine is a measure of accounting comparability proposed by De Franco et al. (2011), which captures the extent to which a firm produces similar financial statements to industry peers in similar economic conditions. Compustat's standardization process may lead to larger discrepancies for firms whose accounting is less comparable to industry peers.

Table 2, Panel A presents the summary statistics of the data discrepancies and the potential factors associated with the discrepancy. On average, the overall discrepancy (*Discr*) across all data items examined is 0.013 (i.e., 1.3% of total assets). The balance sheet, income statement, and cash flow statement have an average discrepancy of 0.016, 0.015, and 0.007, respectively, suggesting that cash flow statement items are least affected by standardization. For an average firm, the level of *Depth* is 5.808; there are 98.9 tags on the three financial statements, of which 9.1% and 6.8% are industry-specific tags and custom tags, respectively.

We then sort firms into deciles by *Discr* and test whether firm characteristics vary across deciles in a predicted way. Panel B of Table 2 reports the means of the characteristics for each *Discr* decile. The average *Depth* almost monotonically increases as we move from decile 1 (smallest *Discr*) to decile 10 (largest *Discr*). This suggests that Compustat's adjustments are greater when the data item is computed from more disaggregated values, a task that could be automated if Compustat relied on the XBRL filing and the associated taxonomy instead of the conventional document formats (e.g., text or html). Filings with a greater data discrepancy also tend to have more tags and higher percentages of industry-specific tags and custom tags. Lastly, greater data discrepancy is associated with lower accounting comparability. These patterns are evident in terms of both the differences between the extreme deciles and the rank correlations. For example, decile 10 has 7.205 ($t = 4.33$) more tags than decile 1, and the rank correlation between *Discr* and *NTag* is 0.278 ($t = 2.60$).

We also examine the cross-industry variations in data discrepancy. Panel C reports the means of data discrepancies for each of the Fama-French 12 industries (excluding financial, utility, and other industries), where industries are sorted by average *Discr*. We find that energy, shops, and durables are the top three industries with the greatest discrepancies, while business equipment and health have relatively small discrepancies.

In Panel D of Table 2, we use multivariate regression analysis to examine the potential determinants of the data discrepancy. In addition to the factors discussed above, we also include the following firm characteristics: logarithm of total assets (*Log(TA)*), logarithm of book-to-market ratio (*Log(BM)*), return on assets (*ROA*), logarithm of the number of analysts covering the firm (*Log(AF)*), return volatility (*RetVol*), logarithm of the number of business segments (*Log(NSeg)*), special items (*SI*), restructuring indicator (*Restructure*), and merger and acquisition indicator (*MA*). Detailed definitions of these variables are provided in Appendix A.

The regression results confirm the findings of Panel B. *Depth*, *Log(NTag)*, and *CustomTag* are positively associated with the discrepancy after controlling for various firm characteristics and fixed effects. Accounting comparability remains negatively associated with data discrepancy. In addition, companies with the largest discrepancies tend to be smaller, less profitable, have more special items, and are less likely to engage in merger and acquisition activities.

4. Properties of accrual accounting

We now turn to the question of whether data discrepancies are consequential in several important research settings. Our analysis starts with research on the properties of accrual accounting, which are central to the ability of earnings to measure firm performance (FASB, 2018; Dechow, 1994). Accruals reduce the temporal fluctuations in cash flows, leading to two well-studied consequences: (i) accruals and cash flows are negatively correlated (Dechow, 1994); and (ii) by including accruals, earnings should be better able to predict future cash flows than current cash flows. Based on these arguments, prior research has developed measures of abnormal accruals and other measures of earnings quality (Dechow et al., 2010). In this section, we examine whether the data discrepancy affects research inferences related to these strands of literature.

4.1. The relationship between accruals and cash flows

The timing role of accrual accounting implies that contemporaneous accruals and cash flows are negatively correlated (Dechow, 1994). In a comprehensive reevaluation of this relationship, Bushman et al. (2016) show this negative correlation has diminished over the years, potentially due to one-time items. We examine whether the negative correlation is affected by the data discrepancies by replicating their analysis over our sample period.

Following Bushman et al. (2016), we estimate the following cross-sectional regressions for each year over 2012–2019:

$$TACC_{i,t} = \beta_0 + \beta_1 CFO_{i,t} + e_{i,t} \tag{1}$$

where *TACC* is total accruals, and *CFO* is cash flow from operations. Theoretically, the coefficient β_1 should be negative. Bushman et al. (2016) show that the coefficient is negative, but the magnitude has been declining over time.

Table 2
Potential factors associated with data discrepancy.

Panel A: Descriptive statistics

	N	Mean	Std. Dev.	Min.	Q1	Median	Q3	Max.
Discr	20,410	0.013	0.017	0	0.005	0.009	0.017	0.751
Discr_BS	20,410	0.016	0.020	0	0.004	0.010	0.020	0.546
Discr_IS	20,410	0.015	0.024	0	0.003	0.006	0.016	0.318
Discr_CF	20,410	0.007	0.015	0	0	0.002	0.006	0.142
Depth	20,410	5.808	0.465	2.623	5.503	5.860	6.132	8.514
NTag	20,410	98.902	19.487	29	86	98	111	202
IndTag	20,410	0.091	0.041	0	0.056	0.086	0.159	0.278
CustomTag	20,410	0.068	0.055	0	0.029	0.055	0.092	0.637
CompAcctInd	7964	−3.106	3.150	−31.832	−3.973	−1.935	−1.093	−0.205
Log(TA)	20,410	5.183	2.964	−6.908	3.453	5.540	7.275	12.836

Panel B: Means of characteristics of *Discr* deciles

	Discr	*Depth*	*NTag*	*IndTag*	*CustomTag*	*CompAcctInd*	*Log(TA)*
Decile 1	0.0017	5.675	89.985	0.090	0.065	−3.050	4.645
2	0.0036	5.771	98.801	0.087	0.056	−2.613	5.943
3	0.0050	5.789	99.779	0.088	0.058	−2.477	5.870
4	0.0065	5.777	100.264	0.088	0.062	−2.659	5.764
5	0.0084	5.808	100.260	0.089	0.062	−2.905	5.658
6	0.0106	5.789	100.561	0.088	0.066	−3.068	5.508
7	0.0134	5.798	100.623	0.090	0.070	−3.268	5.175
8	0.0173	5.798	100.612	0.092	0.074	−3.604	5.042
9	0.0232	5.827	100.764	0.092	0.077	−4.146	4.781
Decile 10	0.0446	5.821	97.190	0.095	0.089	−4.621	3.360
D10−D1		0.145***	7.205***	0.004***	0.024***	−1.571***	−1.285***
(*t-stat.*)		(11.29)	(4.33)	(4.27)	(15.56)	(−4.10)	(−5.42)
Rank Corr.		0.653***	0.278**	0.582***	0.828***	−0.747***	−0.578***
(*t-stat.*)		(10.04)	(2.60)	(8.06)	(23.39)	(−11.32)	(−16.88)

Panel C: Means of characteristics of Fama-French industries

Industry	*Discr*	*Discr_BS*	*Discr_IS*	*Discr_CF*
Business equipment	0.0110	0.0129	0.0123	0.0056
Health	0.0115	0.0116	0.0134	0.0107
Telecommunications	0.0126	0.0157	0.0124	0.0055
Non-durable	0.0128	0.0159	0.0152	0.0045
Manufacturing	0.0130	0.0171	0.0120	0.0043
Chemicals	0.0136	0.0175	0.0161	0.0051
Durable	0.0141	0.0171	0.0150	0.0067
Shops	0.0143	0.0166	0.0207	0.0050
Energy	0.0173	0.0228	0.0152	0.0069

Panel D: Regression analysis

DV = *Discr*	(1)	(2)
Depth	0.211***	0.153**
	(4.98)	(2.56)
Log(NTag)	0.176***	0.120**
	(4.07)	(1.96)
IndTag	0.427	−0.479
	(0.89)	(−0.72)
CustomTag	1.243***	1.403**
	(4.08)	(2.62)
CompAcctInd		−0.021*
		(−1.92)
Log(TA)	−0.043***	−0.028
	(−2.85)	(−1.30)
Log(BM)	−0.067*	−0.119
	(−1.82)	(−1.32)
ROA	−0.218***	−0.225***
	(−9.83)	(−5.87)
Log(AF)	−0.110***	−0.115***
	(−3.96)	(−3.29)
RetVol	0.123	0.355*
	(1.28)	(1.78)
Log(NSeg)	0.038	0.068
	(0.99)	(1.48)
SI	1.113***	0.945***
	(4.90)	(3.61)

(*continued on next page*)

Table 2 (*continued*)

Panel D: Regression analysis

DV = *Discr*	(1)	(2)
Restructure	0.011	0.020
	(0.34)	(0.52)
MA	−0.080***	−0.106***
	(−3.62)	(−3.70)
Industry FE	Yes	Yes
Year FE	Yes	Yes
Adj. R^2	0.203	0.184
N	15,637	7802

Table 2 shows the determinants of the data discrepancy between XBRL and Compustat. In Panel A, we report the descriptive statistics for the following variables: *Discr* is the overall data discrepancy between as-filed and Compustat data, which is calculated as the average of data discrepancies over the 72 accounting variables (excluding total assets) for each firm-year. *Discr_BS*, *Discr_IS*, and *Discr_CF* are the average of data discrepancies for accounting variables from the balance sheet, income statement, and cash flow statement, respectively. In Panel B, we report the mean value of six data discrepancy determinants for each data discrepancy decile. The numbers in each cell are time-series averages of yearly cross-sectional means. We also report the time-series average of the annual rank correlation between *Discr* and each determinant. In Panel C, we report the mean value of data discrepancies for each Fama-French industry. Panel D reports the regression analysis results for the determinants of data discrepancy. In the regressions, in addition to the six determinants, we further control for other firm characteristics. All firm characteristics are measured at the end of the fiscal year. Variable definitions are provided in Appendix A. Significance at the 10%, 5%, and 1% levels (two-sided) are denoted by *, **, and ***, respectively.

Table 3, Panel A reports the results of estimating Equation (1) using both data sources. Based on Compustat, the negative correlation between *TACC* and *CFO* disappeared after 2012, consistent with the finding of Bushman et al. The average coefficient β_1 over our sample period is positive (0.037, $t = 3.47$). However, using as-filed data, we find a negative correlation in all but three years, and the average of β_1 is also negative (−0.088, $t = -1.98$). The difference in β_1 between the two data sources is consistently negative for all years (*Diff_β_1*: −0.125, $t = -2.73$). The explanatory power of the regression (i.e., adjusted R^2) is also greater using as-filed data than using Compustat (*Diff_Adj. R^2*: 0.008, $t = 1.66$).

Further analysis shows that the differences in both the estimated coefficient and the explanatory power are associated with the data discrepancy each year. The rank correlation between *Diff_β_1* and the average discrepancy in the two accounts involved (*Discr_Both*) is −0.810 ($p = 0.01$), suggesting the as-filed accruals-cash flows correlation is more negative than the Compustat accruals-cash flows correlation when the data discrepancy between as-filed and Compustat data is larger. Similarly, the rank correlation between *Diff_Adj. R^2* and *Discr_Both* is 0.762 ($p = 0.03$).

Bushman et al. attribute the decline in the negative association to the increasing prevalence of one-time items caused by complex financial reporting situations. Although we document a continuation of the decline, we also show that the negative correlation is much stronger when using as-filed data. The contrast suggests that Compustat's standardization, at least in part, explains the decline. Therefore, an alternative explanation to Bushman et al.'s findings is that Compustat is more likely to make misclassifications between accruals and cash flows as business transactions have become more complex over the years. In other words, researchers relying on Compustat data will need to separate time trends in the business environment from trends due to Compustat's standardization practices.

Panel B of Table 3 reports the results of replacing the levels of *TACC* and *CFO* with their changes ($\Delta TACC$ and ΔCFO), following Bushman et al. (2016). Even though the coefficient on $\Delta CFO_{i,t}$ is negative using both Compustat and as-filed data, as-filed data yield a more negative coefficient than Compustat (*Diff_β_1*: −0.305, $t = -3.59$), which is closer to the theoretical relationship between the two accounts. The explanatory power using as-filed data is also much greater (*Diff_Adj. R^2*: 0.196, $t = 3.67$). Similar to Panel A, we also show that annual data discrepancies are correlated with the differences in coefficients (−0.762, $p = 0.03$) and explanatory power (0.905, $p < 0.01$).

Over our sample period, there is no discernible temporal trend using Compustat or as-filed data. This finding per se is not inconsistent with Bushman et al. (2016), who show that after 2000, the coefficient is close to −0.5, and there are essentially no temporal trends (see their Figure 1). However, the magnitude of the coefficient using as-filed data is much larger. For example, using change analysis (Table 3, Panel B), the magnitude of the coefficient is larger than 0.7 for most years from 2012 through 2019, which is closer to its theoretical value of −1 than the Compustat counterpart.

Finally, Panel C of Table 3 reports the results of augmenting Equation (1) with the past and future *CFO*, following Dechow and Dichev (2002). The explanatory power of the model has been used to measure the timing role of accruals (i.e., the extent to which accruals curb the temporal fluctuations in the cash flows). We find that the explanatory power using as-filed data is greater than using Compustat (*Diff_Adj. R^2*: 0.105, $t = 2.15$), suggesting that as-filed accruals play a larger timing role than Compustat accruals. Moreover, the difference in explanatory power is positively correlated with the data discrepancy (0.771, $p = 0.07$).

Overall, the results in Table 3 consistently show that Compustat data may understate the negative correlation between accruals and cash flows as well as the explanatory power of the accruals-cash flow regressions. Research findings using Compustat, if interpreted without consideration of data discrepancies, may lead to biased conclusions, especially regarding the temporal trends of the properties of accruals.

Table 3
Contemporaneous relation between total accruals and operating cash flows.

Panel A: Regression model: $TACC_{i,t} = \beta_0 + \beta_1 CFO_{i,t} + \varepsilon_{i,t}$

Year	*As-filed*		*Compustat*		*As-filed vs. Compustat*				
	β_1 *(CFO$_{i,t}$)*	*Adj. R^2*	β_1 *(CFO$_{i,t}$)*	*Adj. R^2*	*Diff_β_1*	*Diff_Adj. R^2*	*Discr_TACC*	*Discr_CFO*	*Discr_Both*
2012	−0.032	0.002	−0.024	0.001	−0.007	0.001	0.010	0.004	0.007
2013	0.068	0.013	0.074	0.016	−0.006	−0.004	0.010	0.005	0.007
2014	−0.167	0.017	0.060	0.011	−0.227	0.006	0.071	0.064	0.067
2015	−0.195	0.019	0.029	0.002	−0.224	0.017	0.089	0.080	0.084
2016	−0.270	0.038	0.031	0.003	−0.301	0.035	0.095	0.090	0.092
2017	−0.167	0.026	0.060	0.014	−0.228	0.012	0.065	0.058	0.061
2018	0.031	0.002	0.041	0.007	−0.010	−0.005	0.018	0.010	0.014
2019	0.026	0.002	0.026	0.003	−0.000	−0.001	0.014	0.005	0.009
Average	−0.088**	0.015***	0.037***	0.007***	−0.125***	0.008*	0.046***	0.039***	0.043***
	(−1.98)	(3.29)	(3.47)	(3.52)	(−2.73)	(1.66)	(3.54)	(3.01)	(3.27)
Trend	0.002	−0.0001	0.002	−0.0003	−0.001	0.0002	0.001	0.0003	0.0004
	(0.08)	(−0.04)	(0.47)	(−0.31)	(−0.03)	(0.10)	(0.11)	(0.05)	(0.08)
Rank Corr.					−0.810**	0.762**			
with *Discr_Both*					[0.01]	[0.03]			

Panel B: Regression model: $\Delta TACC_{i,t} = \beta_0 + \beta_1 \Delta CFO_{i,t} + \varepsilon_{i,t}$

Year	*As-filed*		*Compustat*		*As-filed vs. Compustat*				
	β_1 *(CFO$_{i,t}$)*	*Adj. R^2*	β_1 *(CFO$_{i,t}$)*	*Adj. R^2*	*Diff_β_1*	*Diff_Adj. R^2*	*Discr_ΔTACC*	*Discr_ΔCFO*	*Discr_Both*
2012	−0.544	0.239	−0.418	0.142	−0.125	0.097	0.008	0.004	0.006
2013	−0.347	0.080	−0.404	0.120	0.057	−0.040	0.014	0.006	0.010
2014	−0.722	0.407	−0.430	0.122	−0.292	0.285	0.063	0.054	0.058
2015	−0.867	0.317	−0.277	0.045	−0.590	0.272	0.065	0.055	0.060
2016	−0.898	0.374	−0.422	0.119	−0.476	0.255	0.062	0.054	0.058
2017	−0.854	0.448	−0.345	0.096	−0.509	0.352	0.084	0.077	0.080
2018	−0.843	0.439	−0.389	0.104	−0.454	0.334	0.069	0.057	0.063
2019	−0.371	0.096	−0.324	0.084	−0.047	0.013	0.022	0.011	0.016
Average	−0.681***	0.300***	−0.376***	0.104***	−0.305***	0.196***	0.048***	0.040***	0.044***
	(−8.41)	(5.76)	(−19.37)	(9.86)	(−3.59)	(3.67)	(4.69)	(3.97)	(4.34)
Trend	−0.020	0.012	0.010	−0.006	−0.030	0.017	0.005	0.005	0.005
	(−0.54)	(0.48)	(1.23)	(−1.33)	(−0.80)	(0.72)	(1.19)	(1.04)	(1.12)
Rank Corr.					−0.762**	0.905***			
with *Discr_Both*					[0.03]	[0.00]			

Panel C: Regression model: $TACC_{i,t} = \beta_0 + \beta_1 CFO_{i,t-1} + \beta_2 CFO_{i,t} + \beta_3 CFO_{i,t+1} + \varepsilon_{i,t}$

Year	*As-filed*	*Compustat*	*As-filed vs. Compustat*					
	Adj. R^2	*Adj. R^2*	*Diff_Adj. R^2*	*Discr_TACC*	*Discr_CFO$_{t-1}$*	*Discr_CFO$_t$*	*Discr_CFO$_{t+1}$*	*Discr_All*
2013	0.207	0.265	−0.058	0.010	0.003	0.003	0.038	0.006
2014	0.394	0.175	0.220	0.052	0.003	0.046	0.058	0.058
2015	0.356	0.123	0.232	0.063	0.052	0.052	0.057	0.074
2016	0.205	0.169	0.036	0.062	0.053	0.063	0.039	0.088
2017	0.406	0.209	0.196	0.052	0.067	0.039	0.010	0.062
2018	0.245	0.222	0.023	0.015	0.039	0.008	0.003	0.014
Average	0.302***	0.194***	0.105**	0.042***	0.036***	0.035***	0.034***	0.037***
	(7.90)	(9.68)	(2.15)	(4.35)	(3.24)	(3.55)	(3.62)	(4.94)
Trend	0.002	−0.002	0.004	0.001	0.011**	0.0004	−0.010**	0.001
	(0.08)	(−0.14)	(0.12)	(0.12)	(2.15)	(0.08)	(−2.48)	(0.12)
Rank. Corr.			0.771*					
with *Discr_All*			[0.07]					

Table 3 presents regression results for the contemporaneous relation between total accruals (*TACC*) and operating cash flows (*CFO*) over time for the sample period 2012–2019. In this table, all regressions are estimated annually. In Panel A, β_1 *(CFO$_{i,t}$)* and *Adj. R^2* are the coefficient estimate and the adjusted R^2, respectively, from the level specification of Dechow (1994) model: $TACC_{i,t} = \beta_0 + \beta_1 CFO_{i,t} + \varepsilon_{i,t}$. In Panel B, β_1 *(ΔCFO$_{i,t}$)* and *Adj. R^2* are the coefficient estimate and the adjusted R^2, respectively, from the change specification of Dechow (1994) model: $\Delta TACC_{i,t} = \beta_0 + \beta_1 \Delta CFO_{i,t} + \varepsilon_{i,t}$. In Panel C, *Adj. R^2* is the adjusted R^2 from the Dechow and Dichev (2002) model: $TACC_{i,t} = \beta_0 + \beta_1 CFO_{i,t-1} + \beta_2 CFO_{i,t} + \beta_3 CFO_{i,t+1} + \varepsilon_{i,t}$. In all panels, we also report (i) differences in the coefficient estimate (*Diff_β_1*) and the adjusted R^2 (*Diff_Adj. R^2*) between the as-filed and Compustat models and (ii) data discrepancies in *TACC* (or Δ*TACC*) and *CFO* (or Δ*CFO*). In Panel A (B), *Discr_Both* is the simple average of discrepancies in *TACC* and *CFO* (*ΔTACC* and *ΔCFO*). In Panel C, *Discr_All* is the simple average of discrepancies in *TACC* as well as the last-year (*CFO$_{t-1}$*), current-year (*CFO$_t$*), and next-year operating cash flows (*CFO$_{t+1}$*). Following Bushman et al. (2016), we drop firm-years with average total assets below $10 million. The sample underlying Panels A and B consists of 18,849 firm-year observations with non-missing *TACC* and *CFO* values calculated with both as-filed and Compustat data. We report *t*-statistics in parentheses and *p*-values in brackets. The sample underlying Panel C consists of 10,222 firm-year observations. Variable definitions are provided in Appendix A. Significance at the 10%, 5%, and 1% levels (two-sided) are denoted by *, **, and ***, respectively.

4.2. Cash flow predictability

Central to the timing role of accruals is their ability to predict future cash flows. Prior research has documented conflicting evidence on the relative ability of accruals (and earnings) and cash flows to predict future cash flows (Nallareddy et al., 2020). In this section, we replicate the cash flow predictability tests using Compustat and as-filed data. Specifically, we estimate the following regression:

$$CFO_{i,t} = \beta_0 + \beta^{CFO} CFO_{i,t-1} + \beta^{TACC} TACC_{i,t-1} + \varepsilon_{i,t}. \tag{2}$$

The explanatory power of Equation (2) is denoted $Adj.\ R^2_{CFO,TACC}$. Our analysis focuses on the extent to which accruals possess incremental information beyond cash flows. The incremental explanatory power (predictive ability) of cash flows, *Inc. R^2: CFO*, is defined as $Adj.\ R^2_{CFO,TACC} - Adj.\ R^2_{TACC}$, where $Adj.\ R^2_{TACC}$ is the explanatory power of a univariate regression of *CFO* on lagged *TACC*. Analogously, *Inc. R^2: TACC* is defined as $Adj.\ R^2_{CFO,TACC} - Adj.\ R^2_{CFO}$, where $Adj.\ R^2_{CFO}$ is the explanatory power of a univariate regression of *CFO* on lagged *CFO*.

The results are reported in Table 4, Panel A. Using Compustat data from 2013 to 2019, the incremental predictive ability of cash flows (*Inc. R^2: CFO*) increases during our sample period (*Trend*: 0.007, $t = 2.61$). The incremental predictive ability of accruals (*Inc. R^2: TACC*) generally decreases (*Trend*: −0.001, $t = -1.56$), although the trend is not statistically significant. These findings are consistent with the temporal trends documented by Nallareddy et al. (2020), who attribute their findings to changes in the business environment. However, using as-filed data, we find no such trends (e.g., *Trend* for *Inc. R^2: CFO*: 0.012, $t = 0.40$).

Moreover, different data yield different levels of explanatory power. The incremental predictive ability of accruals is greater using as-filed data than using Compustat data (difference in *Inc. R^2: TACC*: 0.055, $t = 1.77$), whereas the incremental predictive ability of cash flows is smaller using as-filed data than using Compustat data (difference in *Inc. R^2: CFO*: −0.210, $t = -3.85$).

Disaggregating accruals into components could lead to better predictive ability (Barth et al., 2001; Nallareddy et al., 2020). We disaggregate accruals into six major components: change in accounts receivable (*CHG_AR*), change in inventory (*CHG_INV*), change in accounts payable (*CHG_AP*), depreciation expense (*DPExp*), amortization expense (*AMExp*), and net of all other accruals (*Other*). The incremental explanatory power of the six components, when included alongside cash flows, is *Inc. R^2: TACC_Comp*, defined as $Adj.\ R^2_{CFO,TACC\_Comp} - Adj.\ R^2_{CFO}$. Analogously, the incremental explanatory power of cash flows, *Inc. R^2: CFO*, is measured as $Adj.\ R^2_{CFO,TACC\_Comp} - Adj.\ R^2_{TACC\_Comp}$. The results of regression analysis using six accrual components are reported in Panel B of Table 4. Using as-filed data, the incremental contribution of accrual components is significantly higher (0.055, $t = 1.76$), whereas the incremental contribution of cash flows is significantly lower (−0.181, $t = -3.70$) than using Compustat data.

In Panel C of Table 4, we follow Nallareddy et al. (2020) and disaggregate accruals into two groups based on the extent to which accruals are affected by managerial estimates: accruals that are primarily based on managerial estimates (*ACC_EST*) and those largely unaffected by estimates (*ACC_DELTA*). Using Compustat data, we find that *ACC_DELTA* has a greater incremental contribution than *ACC_EST* (0.005 vs. 0.002), consistent with Nallareddy et al. (2020). However, using as-filed data, *ACC_EST* has a greater incremental contribution than *ACC_DELTA* (0.053 vs. 0.005). The incremental contribution of *ACC_EST* is significantly higher using as-filed data than using Compustat (0.051, $t = 1.68$). The incremental contribution of both components of accruals is also higher using as-filed data (0.051, $t = 1.73$).

Collectively, the findings suggest that as-filed accruals and their components (especially accruals based on managerial estimates) exhibit greater incremental predictive ability relative to cash flows than the Compustat counterparts. Moreover, the temporal trend in the declining incremental predictive ability of accruals is likely due to Compustat's standardization practice instead of the attenuation of the timing role of accruals.

4.3. Abnormal accruals

Researchers have developed measures of "abnormal" accruals (or discretionary accruals) based on models of the accrual process (e.g., Dechow and Dichev, 2002). Such measures have been used as a proxy for earnings quality to test predictions on its various determinants and consequences (Dechow et al., 2010).

For each year between 2013 and 2018, we estimate the McNichols' (2002) modification of the Dechow and Dichev (2002) model for each of the Fama-French (1997) 12 industries. We define abnormal accruals (*AbnACC*) as the residual from the cross-sectional regression for each industry:

$$TACC_{i,t} = \beta_0 + \beta_1 CFO_{i,t-1} + \beta_2 CFO_{i,t} + \beta_3 CFO_{i,t+1} + \beta_4 PPE_{i,t+1} + \beta_5 \Delta Rev_{i,t+1} + \varepsilon_{i,t}, \tag{3}$$

where *TACC* is total accruals, *CFO* is cash flows from operations, *PPE* is the net value of property, plant, and equipment, and *ΔRev* is the change in sales revenue.

Table 5, Panel A reports the average coefficients and adjusted R^2. Based on as-filed data, total accruals exhibit more negative (positive) associations with present (past and future) cash flows relative to results using Compustat data. The

Table 4
Cash flow predictability.

Panel A: Total accruals

Year	*Adj.* R^2_{Earn}	*Adj.* $R^2_{CFO,TACC}$	*Adj.* R^2_{CFO}	*Adj.* R^2_{TACC}	*Inc.* R^2*: CFO*	*Inc.* R^2*: TACC*
As-filed						
2013	0.572	0.693	0.682	0.004	0.689	0.012
2014	0.304	0.367	0.363	0.022	0.345	0.004
2015	0.246	0.450	0.426	0.010	0.440	0.024
2016	0.256	0.438	0.396	0.015	0.423	0.042
2017	0.308	0.411	0.286	0.062	0.349	0.126
2018	0.662	0.708	0.466	0.148	0.561	0.243
2019	0.644	0.727	0.705	0.038	0.689	0.021
Average	0.427***	0.542***	0.475***	0.043**	0.499***	0.067**
(*t*-stat.)	(5.98)	(9.02)	(7.87)	(2.24)	(8.93)	(2.04)
Trend	0.035	0.027	0.005	0.014*	0.012	0.022
(*t*-stat.)	(0.99)	(0.87)	(0.15)	(1.77)	(0.40)	(1.42)
Compustat						
2013	0.595	0.702	0.687	0.008	0.694	0.015
2014	0.643	0.734	0.718	0.056	0.678	0.016
2015	0.659	0.757	0.746	0.046	0.711	0.011
2016	0.611	0.734	0.723	0.016	0.718	0.011
2017	0.622	0.732	0.723	0.021	0.711	0.009
2018	0.700	0.790	0.776	0.050	0.740	0.014
2019	0.653	0.747	0.737	0.030	0.717	0.010
Average	0.640***	0.742***	0.730***	0.032***	0.710***	0.012***
(*t*-stat.)	(48.68)	(72.44)	(70.53)	(4.65)	(96.30)	(12.08)
Trend	0.009	0.008*	0.009**	0.001	0.007**	−0.001
(*t*-stat.)	(1.49)	(1.85)	(2.11)	(0.30)	(2.61)	(−1.56)
As-filed vs. Compustat						
Difference	−0.213***	−0.200***	−0.255***	0.011	−0.210***	0.055*
(*t*-stat.)	(−3.13)	(−3.39)	(−4.03)	(0.59)	(−3.85)	(1.77)

Panel B: Disaggregating accruals into six major components

Year	*Adj.* R^2_{CFO}	*Adj.* $R^2_{TACC\_Comp}$	*Adj.* $R^2_{CFO,TACC\_Comp}$	*Inc.* R^2*: TACC_Comp*	*Inc.* R^2*: CFO*
As-filed					
Average	0.475***	0.095***	0.556***	0.082**	0.461***
(*t*-stat.)	(7.87)	(3.95)	(9.28)	(2.44)	(9.07)
Trend	0.005	0.022**	0.026	0.022	0.005
(*t*-stat.)	(0.15)	(2.45)	(0.86)	(1.39)	(0.17)
Compustat					
Average	0.730***	0.115***	0.756***	0.026***	0.642***
(*t*-stat.)	(70.53)	(11.09)	(75.83)	(13.54)	(97.65)
Trend	0.009**	0.008*	0.008**	−0.001	0.0001
(*t*-stat.)	(2.11)	(1.81)	(1.97)	(−0.61)	(0.03)
As-filed vs. Compustat					
Difference	−0.255***	−0.019	−0.200***	0.055*	−0.181***
(*t*-stat.)	(−4.03)	(−0.96)	(−3.40)	(1.76)	(−3.70)

Panel C: Disaggregating accruals based on the magnitude of managerial estimates

Year	*Adj.* R^2_{CFO}	*Adj.* R^2_{CFO,ACC_EST}	*Adj.* R^2_{CFO,ACC_DELTA}	*Adj.* $R^2_{CFO,TACC\_Comp}$	*Inc.* R^2*: ACC_EST*	*Inc.* R^2*: ACC_DELTA*	*Inc.* R^2*: TACC_Comp*
As-filed							
Average	0.475***	0.546***	0.497***	0.550***	0.053*	0.005***	0.076**
(*t*-stat.)	(7.87)	(9.24)	(8.29)	(9.27)	(1.72)	(2.92)	(2.44)
Trend	0.005	0.025	0.006	0.025	0.019	0.0001	0.020
(*t*-stat.)	(0.15)	(0.82)	(0.18)	(0.82)	(1.31)	(0.13)	(1.41)
Compustat							
Average	0.730***	0.749***	0.753***	0.755***	0.002***	0.005***	0.025***
(*t*-stat.)	(70.53)	(74.15)	(72.19)	(73.83)	(4.09)	(7.19)	(26.66)
Trend	0.009**	0.008*	0.009**	0.008*	−0.0004***	0.0004	−0.001
(*t*-stat.)	(2.11)	(1.75)	(2.01)	(1.88)	(−2.76)	(1.23)	(−1.51)
As-filed vs. Compustat							
Difference	−0.255***	−0.204***	−0.256***	−0.205***	0.051*	−0.001	0.051*
(*t*-stat.)	(−4.03)	(−3.49)	(−4.06)	(−3.48)	(1.68)	(−0.36)	(1.73)

Table 4 reports the explanatory power of the regression of current operating cash flows on lagged earnings, operating cash flows, total accruals, and accruals component over time for the sample period 2012–2019. All regressions in this table are estimated annually. In Panel A, *Adj.* R^2_{Earn}, *Adj.* R^2_{CFO}, and *Adj.* R^2_{TACC} are the explanatory power of earnings, operating cash flows, and total accruals-only regression models, respectively. *Adj.* $R^2_{CFO,TACC}$ is the explanatory power of the following regression model:

$CFO_{i,t} = \beta_0 + \beta_1 CFO_{i,t-1} + \beta_2 TACC_{i,t-1} + \varepsilon_{i,t}$.
Inc. R^2: CFO (measured as *Adj.* $R^2_{CFO,TACC}$ − *Adj.* R^2_{TACC}) and *Inc. R^2: TACC* (measured as *Adj.* $R^2_{CFO,TACC}$ − *Adj.* R^2_{CFO}) refer to the incremental explanatory power of cash flows and total accruals, respectively. In Panel B, we decompose total accruals into major components as in Barth et al. (2001). *Adj.* $R^2_{TACC\_Comp}$ (*Adj.* $R^2_{CFO,TACC\_Comp}$) is the explanatory power of the following regression model after excluding (including) lagged operating cash flows as an explanatory variable:
$CFO_{i,t} = \beta_0 + \beta_1 CHG\_AR_{i,t-1} + \beta_2 CHG\_INV_{i,t-1} + \beta_3 CHG\_AP_{i,t-1} + \beta_4 DPExp_{i,t-1} + \beta_5 AMExp_{i,t-1} + \beta_6 Other_{i,t-1} + \beta_7 CFO_{i,t-1} + \varepsilon_{i,t}$.
Inc. R^2: TACC_Comp (measured as *Adj.* $R^2_{CFO,TACC\_Comp}$ − *Adj.* R^2_{CFO}) and *Inc. R^2: CFO* (measured as *Adj.* $R^2_{CFO,TACC\_Comp}$ − *Adj.* $R^2_{TACC\_Comp}$) refer to the incremental explanatory power of total accruals components and cash flows, respectively. In Panel C, we decompose total accruals into components based on managerial estimates. *Adj.* R^2_{CFO,ACC_EST} is the explanatory power of the following regression model:
$CFO_{i,t} = \beta_0 + \beta_1 ACC\_EST_{i,t-1} + \beta_2 CFO_{i,t-1} + \varepsilon_{i,t}$.
Adj. R^2_{CFO,ACC_DELTA} is the explanatory power of the following regression model:
$CFO_{i,t} = \beta_0 + \beta_1 ACC\_DELTA_{i,t-1} + \beta_2 CFO_{i,t-1} + \varepsilon_{i,t}$.
Adj. $R^2_{CFO,TACC\_Comp}$ is the explanatory power of the following regression model:
$CFO_{i,t} = \beta_0 + \beta_1 ACC\_EST_{i,t-1} + \beta_2 ACC\_DELTA_{i,t-1} + \beta_3 CFO_{i,t-1} + \varepsilon_{i,t}$.
Inc. R^2: ACC_EST (measured as *Adj.* R^2_{CFO,ACC_EST} − *Adj.* R^2_{CFO}) refers to the incremental explanatory power of lagged accruals affected by managerial estimates for predicting current cash flows. *Inc. R^2: ACC_DELTA* (measured as *Adj.* R^2_{CFO,ACC_DELTA} − *Adj.* R^2_{CFO}) refers to the incremental explanatory power of lagged accruals unaffected by managerial estimates for predicting current cash flows. *Inc. R^2: TACC_Comp* (measured as *Adj.* $R^2_{CFO,TACC\_Comp}$ − *Adj.* R^2_{CFO}) refers to the incremental explanatory power of lagged accruals components for predicting current cash flows. *Average* is the average explanatory power over the sample period. The sample consists of 16,911 firm-year observations. Variable definitions are provided in Appendix A. Significance at the 10%, 5%, and 1% levels (two-sided) are denoted by *, **, and ***, respectively.

Table 5
Abnormal accruals.

Panel A: Regression results

	As-filed		*Compustat*		*As-filed vs. Compustat*	
	Coef.	*t*-stat.	Coef.	*t*-stat.	Coef.	*t*-stat.
$CFO_{i,t-1}$	0.244***	(6.17)	0.173***	(5.55)	0.071*	(1.93)
$CFO_{i,t}$	−0.595***	(−12.87)	−0.390***	(−11.03)	−0.206***	(−6.08)
$CFO_{i,t+1}$	0.352***	(8.03)	0.270***	(9.34)	0.081**	(2.07)
$PPE_{i,t}$	−0.029***	(−2.73)	−0.056***	(−5.47)	0.027**	(1.97)
$\Delta Rev_{i,t}$	0.030**	(1.99)	0.046***	(3.81)	−0.016	(−1.32)
Adj. R^2	0.334***	(16.13)	0.244***	(13.06)	0.090***	(4.39)

Panel B: Mean characteristics of deciles sorted by *AbnACC* discrepancy

	Discr_AbnACC	*Discr_TACC*	*Discr_CFO$_{t-1}$*	*Discr_CFO$_t$*	*Discr_CFO$_{t+1}$*	*Discr_PPE*	*Discr_ΔRev*	*Discr_All*	*Log(TA)*
Decile 1	0.002	0.009	0.015	0.008	0.012	0.076	0.023	0.027	7.228
2	0.005	0.009	0.014	0.009	0.012	0.083	0.014	0.026	7.169
3	0.009	0.009	0.014	0.009	0.013	0.080	0.014	0.026	7.151
4	0.014	0.012	0.013	0.011	0.019	0.099	0.019	0.032	7.021
5	0.020	0.015	0.019	0.016	0.015	0.102	0.025	0.035	6.814
6	0.028	0.024	0.025	0.023	0.023	0.125	0.020	0.043	6.687
7	0.040	0.029	0.036	0.027	0.028	0.131	0.031	0.051	6.366
8	0.060	0.042	0.045	0.039	0.035	0.140	0.030	0.058	6.100
9	0.095	0.062	0.071	0.059	0.050	0.136	0.038	0.071	5.868
Decile 10	0.232	0.214	0.173	0.188	0.155	0.121	0.050	0.138	4.879
Rank. Corr.		0.943***	0.824***	0.915***	0.859***	0.752***	0.707***	0.954***	−0.917***
(*t-stat.*)		(32.96)	(14.23)	(21.62)	(20.08)	(14.08)	(10.95)	(43.34)	(−28.41)

Table 5 shows abnormal accruals (*AbnACC*) estimated for the sample period 2012–2019. Abnormal accruals are measured based on the following cross-sectional regression for each of Fama and French's (1997) 12 industry groups (excluding financial and utility industries) with at least 10 firms.
$TACC_{i,t} = \beta_0 + \beta_1 CFO_{i,t-1} + \beta_2 CFO_{i,t} + \beta_3 CFO_{i,t+1} + \beta_4 PPE_{i,t+1} + \beta_5 \Delta Rev_{i,t+1} + \varepsilon_{i,t}$,
where *TACC* is total accruals. *CFO* is cash flows from operations. *PPE* is the net value of property, plant, and equipment. *ΔRev* is the change in sales revenue. In Panel A, we report and compare the average estimated coefficients and adjusted-R^2 from the as-filed and Compustat regression. In Panel B, for each abnormal accruals discrepancy (*Discr_AbnACC*) decile, we report the mean values of discrepancies in all variables used to estimate the abnormal accruals and total assets. *Discr_All* is the simple average of discrepancies in all variables used to estimate abnormal accruals. The numbers in each cell are time-series averages of yearly cross-sectional means. Additionally, we report the time-series average of the annual rank correlation between *Discr_AbnACC* and discrepancies in variables used to estimate abnormal accruals along with total assets. The sample consists of 15,313 firm-year observations. Variable definitions are provided in Appendix A. Significance at the 10%, 5%, and 1% levels (two-sided) are denoted by *, **, and ***, respectively.

differences in coefficients are also significant ($CFO_{i,t-1}$: 0.071, $t = 1.93$; $CFO_{i,t}$: −0.206, $t = -6.08$; and $CFO_{i,t+1}$: 0.081, $t = 2.07$). Notably, the average explanatory power of the accruals model is about 37% greater using as-filed data (0.334) than using Compustat data (0.244). The difference is also statistically significant (0.090, $t = 4.39$).

After documenting that data source affects the estimation of the Dechow and Dichev (2002) model, we examine how data discrepancies give rise to discrepancies in estimated abnormal accruals. For each year, we first sort observations into deciles based on the discrepancy in abnormal accruals. We then calculate the means of discrepancies in variables used to estimate the abnormal accruals for each decile. In Panel B of Table 5, we report, for each decile, the time-series average of the yearly cross-sectional means of the discrepancies. We show that the discrepancy in estimated *AbnACC* is positively correlated with the discrepancies in the accounting variables in the accruals model. The rank correlations are high, ranging from 0.707 to 0.943.

We also find that the *AbnACC* discrepancy is negatively correlated with firm size −0.917 ($t = -28.41$). In other words, standardizations have a bigger impact on the abnormal accruals of smaller firms.

The absolute value of abnormal accruals (|*AbnACC*|) has also been used in the literature to capture earnings quality (Dechow et al., 2010). In untabulated results, we find that |*AbnACC*| similarly exhibits discrepancies between the two data sources and that data discrepancies in the variables of the accruals model drive the discrepancies in |*AbnACC*|.

5. Real earnings management

Accounting numbers have also been used to capture real operating activities. In an influential study, Roychowdhury (2006) proposes measures of real earnings management (REM) (i.e., the tendency for managers to manipulate real activities to improve reported earnings numbers). Specifically, he estimates abnormal *CFO*, abnormal discretionary expenses, and abnormal production costs and shows that firm-years with greater capital market pressure exhibit greater levels of real earnings management. In this section, we replicate this analysis using as-filed financial statement data.

Following Roychowdhury (2006), we calculate the three measures of REM using the following regression models:

$$CFO_t = \alpha_0 + \alpha_1 Inv\_TA_{t-1} + \beta_1 Sale_t + \beta_2 \Delta Sale_t + \varepsilon_t, \tag{4}$$

$$DISEXP_t = \alpha_0 + \alpha_1 Inv\_TA_{t-1} + \beta Sale_{t-1} + \varepsilon_t, \tag{5}$$

$$PROD_t = \alpha_0 + \alpha_1 Inv\_TA_{t-1} + \beta_1 Sale_t + \beta_2 \Delta Sale_t + \beta_3 \Delta Sale_{t-1} + \varepsilon_t, \tag{6}$$

where Inv_TA_t is the inverse of total assets, $Sale_t$ is sales revenue, $\Delta Sale_t$ is the change in sales revenue, $DISEXP_t$ is discretionary expenses, and $PROD_t$ is production costs. Detailed variable definitions are provided in Appendix A.[20] The estimated model parameters are reported in Table IA.6. Abnormal *CFO*, abnormal discretionary expenses, and abnormal production costs are calculated as the residual term from the corresponding industry-year model.

To examine whether firms that report earnings just above zero have greater REM activities for the period from 2012 to 2019, we estimate the following regressions:

$$Y_t = \alpha + \beta_1 Size_{t-1} + \beta_2 MTB_{t-1} + \beta_3 NI_t + \beta_4 Suspect\_NI_t + \varepsilon_t, \tag{7}$$

where the dependent variable, Y_t, is one of the three REM measures in year t. The variable of interest is $Suspect\_NI_t$, an indicator variable equal to one if net income scaled by total assets is greater than or equal to zero but less than 0.005 in year t, and zero otherwise.

Following Roychowdhury (2006), we estimate Equation (7) in annual cross-section regressions. Table 6 shows the time-series means of the coefficients along with the Fama-MacBeth t-statistics. Panel A presents the results for abnormal CFO. We find that even though the coefficient on *Suspect_NI* is negative for both as-filed (−0.015, $t = -6.28$) and Compustat data (−0.009, $t = -2.97$), their difference is statistically significant (−0.007, $t = -2.87$).

Panel B presents the results for abnormal discretionary expenses. The coefficient on *Suspect_NI* is negative using Compustat data (−0.006, $t = -1.79$), consistent with the finding of Roychowdhury (2006). However, when we use as-filed data to measure abnormal discretionary expenses, the coefficient becomes positive (0.037, $t = 1.79$). The difference between the two coefficients is also statistically significant (0.043, $t = 1.96$). This contrast suggests that data discrepancies drive a result interpreted as suspect firms having greater abnormal discretionary expenses. Panel C shows that for abnormal production costs, both data sources yield similar findings.

Overall, as-filed data yield significantly different coefficients on *Suspect_NI* for two out of three measures of real earnings management. More importantly, for abnormal discretionary expenses, we obtain the opposite signs for the coefficients. This may be due to the fact that cost of goods sold (COGS) and selling, general, and administrative expense (XSGA) have large discrepancies (Table 1, Panel B). Only 20–25% of observations do not exhibit any significant difference in COGS and XSGA between Compustat and as-filed data (Table 10, Panel B).

6. Accounting-based stock return anomalies

A voluminous literature has examined the relationship between accounting information and future stock returns (Richardson et al., 2010). Findings from this literature may influence the investment decisions of institutional investors and have implications for the informational efficiency of stock prices. In this section, we examine whether data discrepancies affect inferences on the accruals anomaly and 20 other accounting-based anomalies.[21]

[20] Consistent with Roychowdhury (2006), all variables in Equations (4)–(6), regardless of the time subscript, are scaled by total assets in year $t-1$.

[21] The anomaly tests could be low-powered due to the limited sample period. However, two notes are in order. First, we do not have an ex-ante prediction of whether a particular anomaly still exists; nor is it our research question. Our goal is to understand whether research inference is contingent on the data sources and to inform readers about the usability of as-filed data and the circumstances in which as-filed data are more appropriate for the research question. Second, it is exactly in cases where tests lack power or data are scarce that data quality is most important for inference.

Table 6
Real earnings management: Comparison of suspect firm-years with the Rest of the sample.

Panel A: Abnormal cash flows

	As-filed		*Compustat*		*As-filed vs. Compustat*	
	Coef.	*t*-stat.	Coef.	*t*-stat.	Coef.	*t*-stat.
Intercept	−0.001***	(−3.68)	−0.001**	(−2.00)	−0.0004	(−0.57)
Size	0.002***	(2.93)	0.0001	(0.24)	0.002**	(2.55)
MTB	0.001*	(1.68)	0.001**	(1.97)	−0.00004	(−0.06)
NI	0.255***	(4.69)	0.397***	(42.71)	−0.143***	(−2.99)
Suspect_NI	−0.015***	(−6.28)	−0.009***	(−2.97)	−0.007***	(−2.87)
Adj. R^2	0.194***	(2.82)	0.378***	(98.97)	−0.184***	(−3.15)

Panel B: Abnormal discretionary expenses

	As-filed		*Compustat*		*As-filed vs. Compustat*	
	Coef.	*t*-stat.	Coef.	*t*-stat.	Coef.	*t*-stat.
Intercept	0.001	(1.00)	−0.001	(−0.46)	0.001	(0.87)
Size	0.002***	(5.51)	0.003***	(3.48)	−0.001	(−0.78)
MTB	0.010***	(13.53)	0.010***	(15.36)	−0.0002	(−0.17)
NI	−0.465***	(−54.80)	−0.437***	(−74.66)	−0.028**	(−2.31)
Suspect_NI	0.037*	(1.79)	−0.006*	(−1.79)	0.043**	(1.96)
Adj. R^2	0.164***	(15.19)	0.172***	(80.78)	−0.008	(−0.82)

Panel C: Abnormal production costs

	As-filed		*Compustat*		*As-filed vs. Compustat*	
	Coef.	*t*-stat.	Coef.	*t*-stat.	Coef.	*t*-stat.
Intercept	−0.003**	(−2.77)	0.0002	(0.28)	−0.003	(−1.34)
Size	0.001	(1.58)	0.004***	(5.77)	−0.003*	(−1.67)
MTB	−0.004***	(−7.56)	−0.005***	(−4.97)	0.001	(0.13)
NI	−0.147***	(−7.40)	−0.255***	(−10.44)	0.108***	(3.44)
Suspect_NI	0.028	(1.22)	0.017*	(1.82)	0.010	(0.65)
Adj. R^2	0.020***	(6.11)	0.071***	(7.61)	−0.051***	(−5.27)

Table 6 examines whether suspect firms have greater real earnings management for the period 2012–2019. We report the results of the following Fama-MacBeth regressions:

$Y_t = \alpha + \beta_1 Size_{t-1} + \beta_2 MTB_{t-1} + \beta_3 NI_t + \beta_4 Suspect\_NI_t + \varepsilon_t$

All regressors are expressed as deviations from the respective (two-digit) industry-year means. The real earnings management activity, *Y*, is proxied by abnormal cash flows, abnormal discretionary expenses, and abnormal production costs in Panels A, B, and C, respectively. The three measures of real earnings management are estimated with the following models, respectively:

$CFO_t = \alpha_0 + \alpha_1 Inv\_TA_{t-1} + \beta_1 Sale_t + \beta_2 \Delta Sale_t + \varepsilon_t$

$DISEXP_t = \alpha_0 + \alpha_1 Inv\_TA_{t-1} + \beta Sale_{t-1} + \varepsilon_t$

$PROD_t = \alpha_0 + \alpha_1 Inv\_TA_{t-1} + \beta_1 Sale_t + \beta_2 \Delta Sale_t + \beta_3 \Delta Sale_{t-1} + \varepsilon_t$

There are 13,575 (12,851, 11,305) firm-year observations in Panel A (B, C). The estimation results are reported in Table IA.6. Following Roychowdhury (1996), *t*-statistics are calculated using standard errors corrected for autocorrelation using the Newey–West procedure. Variable definitions are provided in Appendix A. Significance at the 10%, 5%, and 1% levels (two-sided) are denoted by *, **, and ***, respectively.

6.1. The accruals anomaly

Sloan (1996) finds that firms with relatively high (low) levels of operating accruals (*OPACC*) experience negative (positive) future abnormal stock returns. This finding, known as the accruals anomaly, has been one of the most closely studied regularities in investment research.

We first examine whether as-filed operating accruals are associated with future stock returns through portfolio analysis. On June 30 of each year *t* from 2013 to 2019, we sort stocks into quintiles based on *OPACC* computed from either Compustat or as-filed data for the fiscal year ending in calendar year $t-1$. Quintile 1 (5) denotes the bottom (top) quintile. Monthly value-weighted returns for stocks in those quintiles are calculated from July of year *t* to June of year $t+1$, and the quintile portfolios are rebalanced in June of year $t+1$. We adopt four measures of portfolio returns: excess returns (*Eret*), Fama-French three-factor alphas (*FF3 Alpha*), Carhart (1997) four-factor alphas (*FF4 Alpha*), and Fama-French five-factor alphas (*FF5 Alpha*).

Table 7, Panel A reports the hedge portfolio returns (i.e., the monthly return to the hedge portfolio that takes a long (short) position in the bottom (top) *OPACC* quintile). For Compustat data, the hedge return is not significantly different from zero regardless of the return measure, consistent with prior findings of the gradual attenuation of the accruals anomaly (Green et al., 2011). For example, the excess return to the Compustat hedge portfolio is 0.296% per month ($t = 1.25$). In contrast, the excess return to the hedge portfolio formed based on as-filed accruals is 0.824% per month ($t = 2.86$), more than twice the Compustat raw return. The difference in hedge return is also significant (0.528%, $t = 3.00$). By measuring portfolio returns using factor model alphas, we also find significant differences in hedge return between the two data sources: 0.433%, 0.420%, and 0.437% for the three-, four-, and five-factor models, respectively.

Table 7
Accruals anomaly: hedge portfolio analysis.

Panel A: Hedge portfolio returns

Return measure	As-filed *OPACC*				Compustat *OPACC*				Diff. in Hedge	*t*-stat.	χ^2
	Q1	Q5	Hedge	*t*-stat.	Q1	Q5	Hedge	*t*-stat.			
Eret	1.592	0.768	0.824***	2.86	1.361	1.065	0.296	1.25	0.528***	3.00	
FF3 Alpha	0.275	−0.304	0.579**	2.01	0.145	−0.001	0.146	0.58	0.433***		8.38
FF4 Alpha	0.288	−0.300	0.589**	2.03	0.165	−0.004	0.169	0.68	0.420***		8.06
FF5 Alpha	0.304	−0.306	0.611**	2.14	0.182	0.008	0.174	0.72	0.437***		8.97

Panel B: Overlap in accruals quintile ranks

	As-filed Q1	As-filed Q2	As-filed Q3	As-filed Q4	As-filed Q5
Compustat Q1	15.03%	2.14%	0.87%	0.78%	1.16%
Compustat Q2	2.42%	12.90%	2.47%	1.36%	0.87%
Compustat Q3	0.67%	3.35%	12.12%	2.69%	1.17%
Compustat Q4	0.70%	1.02%	3.61%	12.30%	2.39%
Compustat Q5	1.17%	0.61%	0.93%	2.89%	14.40%

Panel C: Transition probabilities

Year	Probability of firms staying in As-filed Q1	Probability of firms staying in As-filed Q5	Probability of firms staying in Compustat Q1	Probability of firms staying in Compustat Q5
2013	46.28%	38.79%	49.80%	38.63%
2014	46.92%	39.69%	46.32%	39.80%
2015	45.84%	37.50%	46.32%	39.39%
2016	46.38%	36.55%	46.46%	37.38%
2017	43.75%	36.72%	44.24%	38.54%
2018	43.95%	35.36%	45.93%	37.81%
2019	43.61%	36.01%	46.15%	40.43%
All	45.26%	37.21%	46.46%	38.85%

Panel D: Hedge portfolio returns: disagreement sample and agreement sample

	Eret	*t*-stat.	*FF3 Alpha*	*t*-stat.	*FF4 Alpha*	*t*-stat.	*FF5 Alpha*	*t*-stat.
Compustat agrees with As-filed data:								
Long (Compustat Q1, As-filed Q1)	1.524***	2.82	0.185	0.76	0.203	0.83	0.221	0.94
Short (Compustat Q5, As-filed Q5)	0.777	1.50	−0.301	−1.58	−0.298	−1.56	−0.304	−1.61
Hedge	0.747**	2.37	0.486	1.49	0.501	1.53	0.525*	1.71
Compustat disagrees with As-filed data:								
Long (Compustat Q1, As-filed Q2−Q5)	0.802	1.57	−0.138	−0.57	−0.099	−0.43	−0.122	−0.53
Short (Compustat Q5, As-filed Q1−Q4)	1.534***	3.56	0.459**	2.50	0.446**	2.43	0.485***	2.71
Hedge	−0.732**	−2.36	−0.597*	−1.90	−0.546*	−1.82	−0.607**	−1.96
As-filed data disagrees with Compustat:								
Long (As-filed Q1, Compustat Q2−Q5)	1.665***	3.15	0.384	1.39	0.385	1.38	0.414	1.57
Short (As-filed Q5, Compustat Q2−Q5)	0.811	1.61	−0.262	−1.32	−0.253	−1.27	−0.265	−1.32
Hedge	0.854**	2.51	0.647*	1.90	0.639*	1.86	0.679**	2.08

Table 7 examines the accruals anomaly with hedge portfolio analysis. On June 30 of each year t from 2013 to 2019, we sort stocks into quintiles based on operating accruals (*OPACC*), computed from either as-filed or Compustat data, for the fiscal year ending in calendar year $t-1$ scaled by total assets for the fiscal year ending in $t-2$. Monthly value-weighted returns for stock in those quintiles are calculated from July of year t to June of year $t+1$, and the quintile portfolios are rebalanced in June of $t+1$. Panel A reports the average monthly abnormal returns of the hedging portfolios and their t-statistics. Panel B reports the percentage of firms in intersections between Compustat-based operating accruals quintiles and as-filed operating accruals quintiles. Panel C reports the percentage of firms in the bottom and top quintiles of Compustat and as-filed portfolios, respectively, that are also in that portfolio in the subsequent year. Panel D reports the average monthly abnormal returns of alternative hedging strategies. The abnormal returns are calculated using the Fama-French three-factor model (*FF3 Alpha*), Carhart momentum factor model (*FF4 Alpha*), and Fama-French five-factor model (*FF5 Alpha*). Significance at the 10%, 5%, and 1% levels (two-sided) are denoted by *, **, and ***, respectively.

To interpret the difference in the hedge returns, we examine (i) the extent to which the two data sources overlap with each other in terms of quintile groupings sorted by *OPACC* (Panel B of Table 7) and (ii) whether the two data sources generate different levels of shuffling from one period to the next, among observations (Panel C of Table 7). Panel B shows that for about 67% of all firm-years, the two data sources place the observations in the same quintile: 15.03%, 12.90%, 12.12%, 12.30%, and 14.40% of all firm-years are in the Q1−Q5 portfolios identified by both Compustat and as-filed data. Panel C shows that the probabilities that a given stock remains in the bottom- or top-quintile portfolio from one year to the next are quite similar across portfolios formed using Compustat and as-filed data, indicating that neither data source implies significantly more turnover in portfolio composition than the other.

In plain terms, it is the observations whose quintile assignments differ by data source (about 33% of all observations) that drive the difference in hedge returns. Therefore, to control for the overlap in Compustat and as-filed portfolios, we construct a hedge portfolio in which Compustat "disagrees" with as-filed data for the classification of extreme quintiles, in the spirit of control hedge portfolio tests (Hong et al., 2000). Specifically, we sort stocks independently based on the two *OPACC* measures.

For example, “as-filed Q1” denotes the bottom quintile sorted by the firm's as-filed *OPACC* on the June 30 portfolio formation date. We then take a long position in stocks that belong to Compustat Q1 but not to as-filed Q1 and a short position in stocks that belong to Compustat Q5 but not to as-filed Q5. Analogously, we study the cases in which as-filed disagrees with Compustat and form a portfolio by taking a long position in stocks that belong to as-filed Q1 but not to Compustat Q1 and a short position in stocks that belong to as-filed Q5 but not to Compustat Q5.

Table 7, Panel D reports the results. When we form portfolios based on the overlapping portion of both strategies (i.e., when Compustat agrees with as-filed), we find a positive and significant hedge return measured by *Eret* (0.747%, $t = 2.37$) and five-factor alpha (0.525%, $t = 1.71$), but an insignificant hedge return measured by three- and four-factor alphas. When Compustat disagrees with as-filed data, Compustat data generate a negative hedge portfolio return to the accruals strategy (*Eret*: −0.732%, $t = -2.36$), inconsistent with the accruals anomaly. When as-filed disagrees with Compustat data, as-filed data generate a positive hedge return (*Eret*: 0.854%, $t = 2.51$). These results confirm that the positive hedge return observed in the as-filed data is largely driven by the disagreement portion of the sample.

We also conduct Fama-MacBeth cross-sectional regressions to examine whether as-filed accruals predict future returns after controlling for other variables known to explain future returns. The results reported in Table IA.8 show that (i) when separately included in the regression, as-filed (Compustat-based) operating accruals are negatively (not significantly) associated with future returns; and (ii) when both measures of operating accruals are included in the regression, the as-filed (Compustat-based) measure is negatively (positively) associated with future returns. These findings corroborate the portfolio analysis results.

6.2. *Other accounting-based anomalies*

The accruals anomaly is just one of many accounting-based anomalies documented in prior studies (Richardson et al., 2010). We also examine whether data discrepancies matter for other anomalies. We select 20 accounting-based anomalies from Green et al. (2017) and Hou et al. (2020) that satisfy the following criteria: (i) the study that discovered the anomaly is published in a major accounting or finance journal and (ii) the return predictor is constructed with annual frequency accounting variables.

The list of 20 other accounting-based return predictors includes abnormal operating accruals (*AbnOPACC*), asset growth (*AGR*), book-to-market ratio (*BM*), earnings (before depreciation and extraordinary items) to debt ratio (*CashDebt*), cash flows to price ratio (*CFP*), cash-based operating profitability (*CbOP*), current ratio (*Current*), depreciation to plant assets (*Depr*), changes in PPE and inventory (*dPIA*), earnings-to-price ratio (*EP*), gross profitability (*GMA*), growth in long-term net operating assets (*GrLtNOA*), inventory growth (*GrInv*), investment growth (*GrInvest*), leverage (*Lev*), net operating assets (*NOA*), operating profitability (*OP*), quick ratio (*Quick*), real estate (*RealEstate*), and taxable income (*TB*). Detailed definitions of the predictors are provided in Appendix A. The mapping between Compustat items and XBRL tags involved in calculating each variable is provided in Table IA.7. For each return predictor, we focus on whether there is a significant difference in the hedge returns between the data sources.

The results of this analysis are summarized in Table 8, Panel A. The hedge portfolio excess returns for predictors *AbnOPACC, CashDebt*, *GrLtNOA*, *OP*, and *TB* are significantly different between the two data sources. The results of the portfolio analysis for these five predictors are reported in Table IA.9. For *AbnOPACC, CashDebt, GrLtNOA*, and *TB*, the anomaly finding is stronger using as-filed data than using Compustat data.

Building on insights from Section 3.2, we conjecture that the difference in anomaly findings is greater when the return-predictive signal is based on financial statement items that are more disaggregated and deeper in the reporting taxonomy. To test this conjecture, for each return-predictive variable, we define three variables to capture the complexity of Compustat's task when preparing the underlying data items: *# Tags* is the number of tags used to construct the return-predictive variable; *Mean Depth* is the average level of depth among XBRL tags used to construct the variable; and *Max Depth* is the greatest depth of any tag used to construct the variable.

In Panel B of Table 8, the average *Mean Depth* for return-predictive variables with a significant difference is 5.21, compared to 3.99 for the other variables. The difference, 1.22, is also statistically significant. Similarly, the *Max Depth* for variables with a significant difference is significantly greater than that for other variables (8.67 vs. 5.13). These results support the notion that greater complexity of the data collection task leads to consequential Compustat discrepancies.

6.3. *Additional analysis*

6.3.1. *Unrestated Compustat*

On occasion, Compustat restates financial statement data after registrants amend their 10-K and 10-Q filings (Livnat and López-Espinosa, 2008). Even though most academic research on anomalies uses regular Compustat data, a small number of studies have used either unrestated or point-in-time Compustat data (e.g., Green et al., 2011). It is possible that the discrepancies are caused by these amendments rather than by Compustat's standardization practices. To mitigate this concern,

Table 8
Accounting-based anomalies.

Panel A: Summary of anomaly findings

Predictor	*Mean depth*	*Max depth*	*# Tags*	Compustat anomaly	As-filed anomaly	Difference in anomaly
OPACC	4.00	6.00	3.00	No	Yes	Yes
AbnOPACC	4.25	6.00	5.00	No	Yes	Yes
AGR	1.00	1.00	1.00	No	No	No
BM	3.67	4.00	3.00	Yes	Yes	No
CashDebt	5.33	11.00	3.00	No	Yes	Yes
CbOP	6.83	11.00	6.00	Yes	Yes	No
CFP	5.00	6.00	2.00	Yes	Yes	No
Current	2.50	3.00	2.00	No	No	No
dPIA	3.00	4.00	3.00	No	No	No
Depr	7.00	11.00	2.00	No	No	No
EP	4.00	4.00	1.00	No	No	No
GMA	5.00	9.00	2.00	Yes	Yes	No
GrLtNOA	4.20	11.00	5.00	No	Yes	Yes
GrInv	4.00	4.00	1.00	No	No	No
GrInvest	5.00	5.00	1.00	No	No	No
Lev	2.00	2.00	1.00	No	No	No
NOA	3.00	4.00	6.00	Yes	Yes	No
OP	7.00	11.00	3.00	Yes	Yes	Yes
Quick	3.00	4.00	3.00	No	No	No
RealEstate	4.83	5.00	6.00	Yes	Yes	No
TB	6.50	7.00	2.00	Yes	Yes	Yes

Panel B: Difference in anomaly findings and XBRL tag attributes

Difference in anomaly	*Mean depth*	*Max depth*	*# Tags*
Yes	5.21	8.67	3.50
No	3.99	5.13	2.67
Yes − no	1.22*	3.53**	0.83
(*t-stat.*)	(1.80)	(2.71)	(1.20)

Table 8 summarizes the results of analyses on whether inferences drawn about other accounting-based anomalies are affected by the data source. In Panel A, we report three XBRL tag attributes for each anomaly variable. *Mean Depth* is the average level of XBRL tags that we used to construct each anomaly variable. *Max Depth* is the deepest level among tags used to construct the anomaly variable, and *# Tags* is the number of tags we used to construct each anomaly variable. Additionally, we report whether the hedge portfolio return is significant in our sample period using either Compustat data or as-filed data. Finally, we report whether the difference between the Compustat and as-filed hedge *portfolio* returns is significant. We partition the anomaly variables into two groups depending on whether there is a significant difference in the hedge portfolio returns. In Panel B, we report the average of three XBRL tag attributes for both groups. Variable definitions are provided in Appendix A. Significance at the 10%, 5%, and 1% levels (two-sided) are denoted by *, **, and ***, respectively.

we repeat our analysis of the accruals anomaly by replacing the regular (i.e., “restated”) Compustat data with unrestated Compustat data.[22] The results, reported in Table IA.10, are qualitatively the same as our baseline analysis, suggesting that the amendments do not explain our findings.

6.3.2. *Potential data quality issues with XBRL filings*

XBRL tags may contain errors or inconsistencies (Hoitash et al., 2021). These issues may be particularly relevant for analyses of accounting-based anomalies, which tend to involve granular financial statement accounts. We conduct a robustness check using the accruals anomaly as an example. To address the concern that our anomaly findings may be driven by data quality issues in as-filed data rather than by Compustat standardizations, we exclude firm-years that contain an *OPACC*-related error in violation of the data quality rules developed by XBRL US.[23]

The results are reported in Table IA.11. We find that after excluding filings with errors in *OPACC*-related tags, the findings are qualitatively the same as our baseline results. Therefore, our accruals anomaly results are not driven by data quality issues.

7. Disclosure quality

In the preceding analysis, we re-examine empirical research using the *values* of financial statement items. In this section, we examine the hierarchical structure of the financial statement data. Chen et al. (2015) construct a measure of DQ, which

[22] Point-in-time data values are either the unrestated values or the regular Compustat values. Having tested the two cases at both ends of the spectrum, we expect that using the point-in-time data would yield the same conclusion.

[23] The data are retrieved from https://xbrl.us/data-quality/filing-results/. We obtain the entire set of violations through an application programming interface (API). Each violation, depending on its severity, is classified by XBRL US as an “error,” a “warning,” or as “information.” Examples of errors include elements with negative values when the value should be positive. These errors may account for some of the discrepancies between the two data sources. We focus on “errors.” To pinpoint violations that are errors, we use two sets of data integrity tests: the Data Quality Committee ruleset and the xbrlus-cc consistency checks. We parse these error messages to identify the tags involved in the violation.

captures the level of disaggregation of accounting data through a count of non-missing Compustat line items and reflects the extent of details in firms' annual reports. The DQ measure proposed by Chen et al., albeit intuitive and well-motivated, is constrained by the limited number of data items reported by Compustat and is influenced by its standardization practices.

In this section, we follow the spirit of Chen et al. to develop an as-filed measure of DQ based on XBRL filings. Our method has two advantages: (i) instead of relying upon Compustat's balancing model, we use the authoritative FASB taxonomy to quantify the hierarchical structure of each financial statement; and (ii) we use XBRL tags as reported in the original filings, instead of data items that are extracted and standardized by Compustat. Analogous to Chen et al.'s DQ measure (referred to as $DQ^{Compustat}$), our measure ($DQ^{As\text{-}filed}$) captures the percentage of non-missing items on the balance sheet and income statement. The detailed procedure for constructing the measure is described in Appendix C.3. We calculate $DQ^{As\text{-}filed}$ for 8327 firm-year observations from 2012 to 2019, for which $DQ^{Compustat}$ is non-missing as well.

Table 9, Panel A presents the descriptive statistics for the two DQ measures. The mean of $DQ^{Compustat}$ is 0.782, whereas the mean of $DQ^{As\text{-}filed}$ is 0.521. In Panel B, we examine the relationship between as-filed DQ and firm characteristics. Not surprisingly, the two DQ measures are positively correlated (column 1: 0.344, $t = 3.46$).[24] The correlation remains after controlling for firm characteristics that may influence DQ (column 2: 0.351, $t = 3.70$).

We then follow Chen et al. (2015) and conduct a validation test using analyst forecasts. Chen et al. predict and confirm in their sample that higher DQ is associated with lower analyst forecast dispersion and higher analyst forecast accuracy. The main dependent variables in the regressions are $DISP_{i,t+1}$ and $|FE|_{i,t+1}$. $DISP_{i,t+1}$ is the analyst forecast dispersion at year $t+1$, measured as the average of the standard deviation of analyst forecasts for year $t+1$ earnings sampled at each month over year $t+1$. $|FE|_{i,t+1}$ is the average of the mean absolute forecast error for year $t+1$ earnings sampled at each month of year $t+1$.

Panel C presents the regression results. For forecast dispersion, although the coefficient on $DQ^{Compustat}$ is negative, it is not significant during our sample period. However, $DQ^{As\text{-}filed}$ is negative and significant (column 2: -6.193, $t = -2.21$). For forecast accuracy, we find that the coefficient of neither $DQ^{Compustat}$ nor $DQ^{As\text{-}filed}$ is significant, potentially due to lack of power.

Overall, our results suggest that there are both commonalities and differences between the two measures. The differences arise partly from the more comprehensive hierarchical structure of FASB's taxonomy relative to the Compustat balancing model.

8. Financial statement data from FactSet

8.1. Discrepancies between FactSet and as-filed data

In this section, we replicate the analysis by replacing Compustat with financial statement data from FactSet, a major competitor of S&P Global. Of the 73 Compustat data items reported in Table 1, we are able to match 70 to FactSet data items.[25] The descriptive statistics of the data discrepancies between FactSet and as-filed data are shown in Table IA.12. For 67 (95.7%) of the 70 FactSet data items, there are statistically significant discrepancies with the as-filed data. This percentage is comparable with that based on Compustat data.

In Table 10, we contrast the discrepancies between FactSet and as-filed data with those between Compustat and as-filed data. For all but four items, the discrepancies are positively correlated. We then analyze the extent to which each data source “overstates” or “understates” a data item in the overlapping sample.[26] Three patterns emerge. First, for 47 (44) out of 70 data items, Compustat (FactSet) exhibits no difference from as-filed data for more than 80% of the observations. In other words, both data sets are consistent with as-filed data over more than 80% of the sample for most data items. Second, for 41 (46) data items, FactSet overstates (understates) more often than Compustat. On the other hand, for 19 (16) data items, Compustat overstates (understates) more often than FactSet. Therefore, FactSet seems to deviate more often from the as-filed values. Third, lower-level (more granular) items (e.g., TXACH) seem to be more likely to have different values between different data sources.

8.2. Specific research settings

We then examine how discrepancies between FactSet and as-filed data affect inferences in specific research settings. Regarding the properties of accrual accounting, we find that (i) the negative association between accruals and cash flows is stronger in as-filed data than in FactSet data, and the regression model has greater explanatory power when we use as-filed data (Table IA.13); (ii) as-filed accruals possess greater incremental predictive power than the FactSet counterpart; however, different from our findings for Compustat, the incremental predictability of the managerial estimates-driven accruals is not

[24] Following Chen et al. (2015), all coefficients in Table 9 are multiplied by 100 for expositional convenience.

[25] The detailed mapping between the two data sources is provided in Table IA.5 in the Internet Appendix.

[26] For the same accounting variable, if a Compustat or FactSet value differs from as-filed value by less than 0.1%, then we consider there to be no difference; if a Compustat or FactSet value is higher (lower) by more than 0.1%, then we consider the value to be “overstated” (“understated”) relative to as-filed data.

Table 9
Disclosure quality.

Panel A: Descriptive statistics of DQ measures

	N	Mean	Std. Dev.	Min.	Q1	Median	Q3	Max.
$DQ^{Compustat}$	8327	0.782	0.062	0.405	0.755	0.782	0.821	0.924
$DQ^{As\text{-}filed}$	8327	0.521	0.136	0.072	0.427	0.517	0.612	0.879

Panel B: Relationships between the two DQ measures

$DV = DQ^{As\text{-}filed}$	(1)	(2)
$DQ^{Compustat}$	34.372***	35.060***
	(3.46)	(3.70)
Restructure		−1.405**
		(−2.61)
MA		0.317
		(0.82)
SI		−0.001
		(−0.21)
RetVol		2.233
		(0.87)
Log(TA)		−0.733**
		(−2.58)
Log(NSeg)		−0.792
		(−1.33)
Year FE	Yes	Yes
Industry FE	Yes	Yes
N	8327	8327
Adj. R^2	0.129	0.146

Panel C: DQ and analyst forecast properties

DV	*Disp*		*|FE|*	
	(1)	(2)	(3)	(4)
Intercept	36.595*	36.266*	85.488	85.061
	(1.81)	(1.85)	(1.54)	(1.55)
$DQ^{Compustat}$	−34.606	−32.332	−40.442	−37.500
	(−1.62)	(−1.46)	(−1.00)	(−0.87)
$DQ^{As\text{-}filed}$		−6.193**		−8.011
		(−2.21)		(−0.70)
Log(NTag)	6.391*	6.780**	15.128*	15.631*
	(1.83)	(2.12)	(1.68)	(1.87)
σ(EPS)	2.826***	2.819***	6.856***	6.847***
	(9.94)	(10.00)	(8.80)	(8.86)
Growth	3.621**	3.639**	7.607	7.630
	(2.88)	(2.87)	(1.55)	(1.54)
ROA	21.977**	21.804**	58.071**	57.846**
	(2.45)	(2.40)	(2.83)	(2.79)
Log(AF)	3.024*	3.143**	−4.464	−4.311
	(1.86)	(1.96)	(−0.91)	(−0.91)
Log(TA)	−6.004***	−6.093***	−15.584***	−15.699***
	(−6.15)	(−6.37)	(−6.42)	(−6.70)
Firm fundamentals	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
Industry FE	Yes	Yes	Yes	Yes
N	8327	8327	8327	8327
Adj. R^2	0.091	0.091	0.077	0.077

Table 9 examines the disclosure quality scores constructed using Compustat ($DQ^{Compustat}$) and as-filed data ($DQ^{As\text{-}filed}$), separately, for the sample period 2012–2019. The sample in this table includes 8327 firm-year observations with non-missing $DQ^{Compustat}$ and $DQ^{As\text{-}filed}$. Panel A presents descriptive statistics for $DQ^{Compustat}$ and $DQ^{As\text{-}filed}$. Panel B presents the estimation results from the regression of $DQ^{As\text{-}filed}$ on $DQ^{Compustat}$ along with a set of firm fundamentals. Panel C presents the regression analysis for the association between DQ and (i) analyst forecast dispersion (*Disp*) and (ii) the absolute value analyst forecast error (|*FE*|). In both Panels B and C, all coefficients are multiplied by 100 for expositional convenience. Year and two-digit SIC industry fixed effects are included, and standard errors are two-way clustered by year and industry. Variable definitions are provided in Appendix A. Significance at the 10%, 5%, and 1% levels (two-sided) are denoted by *, **, and ***, respectively.

Table 10
Comparison of Compustat and FactSet data discrepancies.

Panel A: Balance sheet

Data Item	*Corr_Diff*	*%FactSet Overstated*	*%Compustat Overstated*	*Diff: % Overstated*	*%FactSet Understated*	*%Compustat Understated*	*Diff: %Understated*	*%FactSet No Difference*	*%Compustat No Difference*	*Diff: %No Difference*
Assets:										
[1] AT										
[2] ACT	0.204***	0.06	0.01	0.05**	6.62	0.00	6.62***	93.33	99.99	−6.67***
[3] CHE	0.486***	9.73	6.63	3.10***	1.85	0.62	1.23***	88.43	92.76	−4.33***
[4] CH	0.070***	17.92	0.30	17.62***	2.36	0.23	2.13***	79.73	99.48	−19.75***
[5] IVST	0.143***	6.29	17.79	−11.49***	2.43	1.09	1.34***	91.28	81.13	10.15***
[6] RECT	0.755***	27.94	26.94	1.00***	1.14	1.19	−0.05	70.92	71.87	−0.95***
[7] INVT	0.203***	4.13	3.21	0.93***	0.46	0.54	−0.07	95.40	96.25	−0.85***
[8] ACO	0.626***	1.18	1.18	0.00	47.93	36.83	11.10***	50.89	61.99	−11.10***
[9] XPP	0.516***	6.83	4.08	2.75***	5.84	4.93	0.90***	87.34	90.99	−3.65***
[10] ACOX	0.587***	5.45	1.97	3.48***	49.48	38.12	11.36	45.07	59.92	−14.84***
[11] ANCT	−0.094***	6.93	0.28	6.65***	0.37	0.29	0.08	92.69	99.49	−6.73***
[12] PPENT	0.995***	39.32	38.98	0.35***	52.09	51.77	0.31***	8.59	9.25	−0.66***
[13] PPEGT	0.996***	39.35	39.09	0.26***	51.38	51.15	0.24***	9.27	9.77	−0.50***
[14] DPACT	0.225***	2.76	3.27	−0.50***	0.27	0.41	−0.15**	96.97	96.32	0.65***
[15] IVAEQ	0.294***	0.00	6.86	−6.86***	9.28	2.06	7.22***	90.72	91.07	−0.36
[16] IVAO	0.110***	1.00	22.73	−21.73***	12.19	1.30	10.88***	86.81	75.97	10.84***
[17] INTAN	0.375***	7.55	7.67	−0.12	3.47	2.57	0.90***	88.98	89.76	−0.78***
[18] INTANO	0.464***	7.69	8.82	−1.12***	3.17	3.11	0.06	89.14	88.07	1.06***
[19] GDWL	0.095***	0.31	0.35	−0.04	0.73	0.36	0.37***	98.96	99.29	−0.33***
[20] AO	0.900***	46.81	47.67	−0.86***	52.53	50.17	2.36***	0.66	2.15	−1.50***
Liabilities and shareholders' equities:										
[21] LT	−0.120***	0.65	0.12	0.53***	0.82	0.11	0.71***	95.53	99.76	−1.24***
[22] LCT	0.053***	0.07	0.03	0.04*	0.11	0.04	0.07**	99.81	99.93	−0.11***
[23] DLC	0.622***	8.22	7.86	0.37**	8.67	6.69	1.98***	83.11	85.45	−2.35***
[24] DD1	0.366***	6.05	10.68	−4.64***	9.46	5.63	3.83***	84.50	83.69	0.81**
[25] NP	0.271***	11.18	3.49	7.70***	2.68	2.08	0.60***	86.14	94.43	−8.30***
[26] AP	0.236***	4.85	4.40	0.45**	3.92	1.15	2.77***	91.24	94.45	−3.22***
[27] TXP	0.351***	10.52	2.03	8.49***	6.20	5.17	1.03***	83.29	92.81	−9.52***
[28] LCO	0.683***	12.24	8.93	3.32***	35.33	26.90	8.43***	52.43	64.18	−11.75***
[29] XACC	0.746***	33.24	27.63	5.62***	39.15	22.83	16.31***	27.61	49.54	−21.93***
[30] LCOX	0.752***	37.95	24.46	13.50***	46.57	41.06	5.51***	15.48	34.48	−19.01***
[31] LNCT	−0.067***	0.71	0.76	−0.05	1.00	0.80	0.20*	98.29	98.44	−0.15
[32] DLTT	0.653***	11.34	11.87	−0.53***	2.56	2.74	−0.18**	86.10	85.39	0.71***
[33] TXDITC	0.018**	1.69	0.18	1.51***	44.18	0.48	43.70***	54.13	99.34	−45.21***
[34] LO	0.659***	30.28	2.77	27.51***	45.15	56.82	−11.67***	24.58	40.41	−15.84***
[35] TEQ	0.173***	0.33	0.11	0.22***	0.91	0.11	0.80***	98.77	99.77	−1.00***
[36] SEQ	0.174***	0.85	0.16	0.69***	6.99	0.64	6.34***	92.17	99.19	−7.03***
[37] CEQ	0.380***	0.73	0.25	0.48***	9.72	3.42	6.30***	89.55	96.33	−6.78***
[38] CSTK	0.055***	3.61	4.70	−1.09***	0.33	4.59	−4.26***	96.06	90.71	5.35***
[39] RE	0.042***	0.38	14.61	−14.23***	0.78	49.55	−48.78***	98.85	35.84	63.01***
[10] PSTK	0.095***	1.04	1.19	−0.15	0.15	0.57	−0.43***	98.81	98.24	0.57***
[41] MIBN	0.843***	5.09	0.05	5.04***	0.19	0.08	0.11**	94.72	99.87	−5.15***

Panel B: Income statement

Data Item	*Corr_Diff*	*%FactSet Overstated*	*%Compustat Overstated*	*Diff: Overstated*	*%FactSet Understated*	*%Compustat Understated*	*Diff: Understated*	*%FactSet No Difference*	*%Compustat No Difference*	*Diff: No Difference*
[42] SALE	0.493***	5.65	5.56	0.09	7.66	5.72	1.95***	86.69	88.72	−2.04***
[43] COGS	0.488***	53.09	12.27	40.82***	19.97	62.76	−42.79***	26.95	24.97	1.97***
[44] XSGA	0.816***	15.65	13.46	2.19***	61.71	65.04	−3.33***	22.64	21.49	1.15***
[45] XRD	0.152***	2.79	1.83	0.96***	7.04	1.06	5.99***	90.17	97.11	−6.95***
[46] DP	0.722***	27.28	25.30	1.98***	6.05	16.62	−10.57***	66.68	58.09	8.59***
[47] AM	0.348***	23.59	9.24	14.35***	4.51	3.24	1.28***	71.90	87.53	−15.63***
[48] OIADP	0.768***	54.73	57.66	−2.92***	14.08	12.38	1.70***	31.19	29.97	1.22***
[49] XINT	0.384***	13.58	20.35	−6.77***	13.38	6.91	6.46***	73.04	72.74	0.30
[50] PI	0.515***	6.66	5.85	0.82***	7.02	4.26	2.76***	86.32	89.89	−3.57***
[51] TXT	0.185***	0.97	1.95	−0.98***	1.22	1.78	−0.55***	97.81	96.27	1.54***
[52] IB	0.260***	2.41	7.35	−4.94***	11.21	5.90	5.31***	86.38	86.76	−0.38

Panel C: Cash flow statement

Variable	*Corr_Diff*	*%FactSet Overstated*	*%Compustat Overstated*	*Diff: Overstated*	*%FactSet Understated*	*%Compustat Understated*	*Diff: Understated*	*%FactSet No Difference*	*%Compustat No Difference*	*Diff: No Difference*
Operating activities:										
[53] OANCF	0.735***	14.00	10.07	3.93***	12.99	10.44	2.55***	73.01	79.50	−6.49***
[54] IBC	0.149***	2.22	5.89	−3.66***	3.42	5.22	−1.80***	94.36	88.90	5.46***
[55] OPCAPCH	0.535***	9.62	10.23	−0.61***	10.54	10.05	0.49*	79.85	79.72	0.12

Table 10 (*continued*)

Panel C: Cash flow statement

Variable	*Corr_Diff*	*%FactSet Overstated*	*%Compustat Overstated*	*Diff: Overstated*	*%FactSet Understated*	*%Compustat Understated*	*Diff: Understated*	*%FactSet No Difference*	*%Compustat No Difference*	*Diff: No Difference*
[56] RECCH	0.543***	7.21	8.19	−0.99***	7.88	9.12	−1.23***	84.91	82.69	2.22***
[57] INVCH	0.316***	4.71	2.93	1.78***	4.50	3.49	1.01***	90.79	93.28	−2.79***
[58] TXACH	0.165***	22.07	1.62	20.45***	16.62	1.45	15.17***	61.304	96.93	−35.62***
[59] APALCH	0.439***	21.61	16.99	4.62***	30.65	16.78	13.88***	47.74	66.23	−18.49***
[60] AOLOCH	0.413***	27.56	36.28	−8.72***	30.45	31.59	−1.139*	41.98	32.13	9.86***
[61] DPC	0.701***	21.82	38.65	−16.83	13.09	7.84	5.26***	65.09	53.51	11.57***
[62] XIDOC	0.047	0.55	2.99	−2.43***	0.00	5.58	−5.58***	99.45	91.44	8.01***
Investing activities:										
[63] IVNCF	0.129***	3.55	1.80	1.75***	5.41	2.12	3.29***	91.04	96.08	−5.04***
[64] CAPX	0.298***	25.68	7.84	17.84***	0.57	2.05	−1.49***	73.76	90.11	−16.35***
[65] SPPE	0.222***	17.77	2.68	15.09***	0.61	1.24	−0.63***	81.63	96.09	−14.46***
[66] AQC	0.459***	6.27	3.72	2.55***	2.53	2.72	−0.19*	91.21	93.56	−2.35***
Financing activities:										
[67] FINCF	0.113***	4.22	0.62	3.60***	4.76	1.54	3.21***	91.03	97.84	−6.82***
[68] SSTK	0.641***	10.18	10.38	−0.20	10.17	6.39	3.78***	79.65	83.23	−3.58***
[69] PRSTKC	0.285***	5.57	18.88	−13.31***	1.76	1.42	0.33*	92.68	79.70	12.98***
[70] DV	0.710***	2.25	2.09	0.15*	4.88	4.86	0.02	92.87	93.05	−0.18

Table 10 compares Compustat- and FactSet-based data discrepancies during the sample period 2012–2019. *Corr_Diff* is the time-series average of the annual rank correlation between *Diff_Compustat* and *Diff_FactSet*, where *Diff_Compustat* (*Diff_FactSet*) is calculated as the signed difference between the Compustat (FactSet) value and the as-filed value of each accounting variable. For each accounting variable, we calculate a ratio of Compustat or FactSet value to as-filed value. If the ratio is within the range [99.9%, 100.1%], then it is deemed as *no difference* between as-filed value and Compustat or FactSet value. If Compustat or FactSet value is higher (lower) by more than 0.1%, then we consider the value to be "overstated" ("understated") relative to as-filed data. *%Compustat Overstated* (*%Compustat Understated*, *%Compustat No Difference*) is the percentage of observations for which the Compustat value is greater than (less than, no different from) the corresponding as-filed value. Analogously, *%FactSet Overstated* (*%FactSet Understated*, *%FactSet No Difference*) is the percentage of observations for which FactSet value is greater than (less than, no different from) the corresponding as-filed value. *Diff: %Overstated* is calculated as *%FactSet Overstated* minus *%Compustat Overstated. Diff: %Understated* is *%FactSet Understated* minus *%Compustat Understated. Diff: %No Difference* is *%FactSet No Difference* minus *%Compustat No Difference*. Significance at the 10%, 5%, and 1% levels (two-sided) are denoted by *, **, and ***, respectively. Indented data items are components of higher-level data items.

statistically different between FactSet and as-filed data (Table IA.14); and (iii) as-filed data fit the regression model better than the FactSet data, and the discrepancies in abnormal accruals tend to be larger for smaller firms (Table IA.15).

We report the replication of real earnings management in Table IA.16. We find that, using both as-filed and FactSet data, suspect firm-years have lower abnormal CFO but do not have different magnitudes of abnormal discretionary expenses and abnormal production costs during our sample period. There are also no significant differences between FactSet and as-filed data for any of the three measures.[27]

Finally, we re-examine accounting-based anomalies using FactSet data. The hedge portfolio returns following the accruals strategy based on the FactSet data are generally smaller and less significant than those based on the as-filed data (Table IA.17). In addition, we observe significant differences between FactSet and as-filed in three other anomalies (i.e., operating profitability, cash-based profitability, and earnings-to-debt ratio; see Tables IA.18 and IA.19).

In summary, FactSet data also exhibit significant discrepancies from as-filed data. These discrepancies are different from and more pervasive than Compustat discrepancies. In common with Compustat discrepancies, FactSet discrepancies are large enough to affect inference in most research settings we have examined.

9. Concluding remarks

We show that more than 90% of the Compustat and FactSet variables we examine contain large and frequent discrepancies from as-filed data. These discrepancies differ by the data provider and are large enough to affect inference. Data discrepancies (i) reduce the fit of regressions of cash flows on accruals, which clouds interpretations of the temporal trends previously studied using Compustat data; (ii) affect inferences on whether suspect firms have more aggressive real earnings management activities; and (iii) shuffle the sorting of companies into hedge portfolios based on return-predictive accounting signals. Furthermore, measures of DQ based on Compustat's spare balancing model and the more comprehensive hierarchy embedded in the FASB taxonomy are correlated but by no means equivalent.

Our quantification of the prevalence and magnitude of discrepancies in commonly used data products as well as their effects on inference should inform users' choice of the data source. Moreover, our publicly available as-filed data facilitate the use of structured disclosures by researchers and the investment community. Further improvements of structured disclosures' reliability (e.g., via enhanced assurance, such as by expanding the audit opinion to include the structured data in filings), granularity (by further refining the taxonomy), and usability (e.g., by creating more capable APIs) merit further exploration.

[27] We note that the sample for the FactSet analysis is smaller than the sample used in the comparison between Compustat and as-filed data.

Our analysis is conducted in the context of the increasing availability of structured disclosures and an evolving data aggregation industry. Technological advances and standards for structured disclosures have made regulatory filings accessible via automated algorithms and shortened the distance between preparers and end-users of financial statements. They also narrow the circumstances in which a user must rely on data aggregators' products assembled using proprietary, and therefore somewhat opaque, standardizations. Those products nevertheless remain useful when long sample periods or comparisons across jurisdictions that do not mandate structured disclosures are required.

Appendix A. Definitions of variables[28]

Variable	Definitions
AbnACC	Abnormal total accruals are the residual terms of the cross-sectional regression: $TACC_{i,t} = \beta_0 + \beta_1 CFO_{i,t-1} + \beta_2 CFO_{i,t} + \beta_3 CFO_{i,t+1} + \beta_4 PPE_{i,t+1} + \beta_5 \Delta Rev_{i,t+1} + \varepsilon_{i,t}$, where *TACC* is total accruals, *CFO* is cash flows from operating activities (OANCF), *PPE* is net value of property, plant, and equipment (PPENT) and *ΔRev* is change in revenue (SALE). All variables are scaled by average total assets. We estimate the above equation for each of Fama and French's (1997) 12 industry groups (excluding financial and utility industries) with at least 10 firms in year *t*.
AbnOPACC	Abnormal operating accruals are measured analogous to *AbnACC*, with the dependent variable (*TACC*) replaced by operating accruals (*OPACC*). All variables are scaled by average total assets.
ACC_DELTA	Accruals largely unaffected by managerial estimates, calculated as RECCH + APALCH + TXACH + AOLOCH.
ACC_EST	Accruals primarily based on managerial estimates, calculated as *TACC* − *ACC_DELTA.*
AF	Number of analysts issuing EPS forecasts for the current year.
AGR	Growth in total assets (AT).
AMExp	Amortization expenses (AM), scaled by average total assets.
Beta	Market beta, estimated from a regression of weekly returns on equal-weighted market returns for the previous three years ending in month t−1 with at least 52 weeks of returns.
BM	Book-to-market ratio, which is calculated as the book value of equity for the fiscal year ending in calendar year $t-1$ divided by the market value of equity on December 31 of year $t-1$. Book value of equity is calculated as stockholders' book equity, plus balance-sheet deferred taxes and investment tax credit (TXDITC) if available, minus the book value of preferred stock. Stockholders' equity is the value reported by Compustat (SEQ), if available. If unavailable, stockholders' equity is the book value of common equity (CEQ) plus the par value of preferred stock (PSTK), or total assets (AT) minus total liabilities (LT). Depending on availability, we use redemption (PSTKRV), liquidating (PSTKL), or par value (PSTK) for the book value of preferred stock.
CashDebt	Earnings before depreciation and extraordinary items-to-debt ratio, defined as the sum of earnings before extraordinary items (IB) and depreciation (DP) divided by average total liabilities (LT).
CbOP	Cash-based operating profitability, defined as operating profitability (OP) + decrease in accounts receivable (RECCH) + decrease in inventory (INVCH) + increase in accounts payable and accrued liabilities (APALCH).
CHG_AP	Changes in accounts payable (APALCH), scaled by average total assets.
CHG_AR	Changes in accounts receivable (RECCH), scaled by average total assets.
CHG_INV	Changes in inventory (INVCH), scaled by average total assets.
CFO	Cash flows from operating activities (OANCF), scaled by average total assets (or lagged total assets in Section 5).
CFP	Cash flows-to-price ratio is defined as cash flows from operating activities (OANCF) divided by the market value of equity (PRCC_F × CSHO).
CompAcctInd	Accounting comparability measure proposed by De Franco et al. (2011). It captures the extent to which two companies produce similar financial statements given the same underlying economic conditions. To compute the accounting comparability between firm *i* and firm *j* in the same three-digit SIC industry in year *t*, we first regress firm *i*'s (firm *j*'s) earnings on returns to obtain the intercept $\widehat{\alpha}_i$ and coefficient $\widehat{\beta}_i$ on returns ($\widehat{\alpha}_j$ and $\widehat{\beta}_j$). Then, we calculate the predicted earnings for firm *i* (firm *j*) using $\widehat{\alpha}_i$ and $\widehat{\beta}_i$ ($\widehat{\alpha}_j$ and $\widehat{\beta}_j$). We then calculate the comparability for each firm pair (*i*, *j*), $CompAcct_{ijt}$, as the negative value of the average absolute difference between the predicted earnings in the past 16 quarters, divided by 100. $CompAcctInd_{i,t}$ is the median of $CompAcct_{ijt}$, where firm *j* is firm *i*'s peer firm in the same three-digit SIC industry in year *t*.
Current	Current ratio, defined as current assets (ACT) divided by current liabilities (LCT).
CustomTag	Percentage of custom tags out of all tags used on three financial statements.
Depr	Depreciation-to-plant asset ratio, defined as depreciation and amortization expenses (DP) divided by net value of property, plant, and equipment (PPENT).
Depth	Average level of depth among all tags used on three financial statements.
Discr	Overall data discrepancy between as-filed and Compustat data, calculated as the average of data discrepancies among all commonly used accounting variables.
DISEXP	Discretionary expenses, calculated as the sum of R&D (XRD), advertising (XAD), and selling, general, and administrative (XSGA) expenses, scaled by lagged total assets. As long as XSGA is available, XRD and XAD are set to zero if they are missing.
Disp	Analyst forecast dispersion, measured as the average standard deviation of analyst forecast of year $t+1$ sampled at each month over year *t*.
DPExp	Depreciation expense, measured as DP − AM, scaled by average total assets. Compustat item DP is depreciation and amortization expenses; AM is the amortization of intangible assets.
dPIA	Change in PP&E and inventory-to-assets, defined as the change in gross property, plants, and equipment (PPEGT) plus the change in inventory (INVT) scaled by lagged total assets.
$DQ^{As\text{-}filed}$	Disclosure quality score based on as-filed data. The construction details are provided in Appendix C.3.
$DQ^{Compustat}$	Disclosure quality based on Compustat data, constructed by Chen et al. (2015).

[28] For accounting-based variables, the definition is based on Compustat data items. The XBRL version is defined using the same formula as its Compustat-based counterpart, but the data items are constructed from XBRL data. The mappings between Compustat data items and XBRL tags involved in calculating these variables are provided in Table IA.4 of the Internet Appendix.

(*continued*)

Variable	Definitions
EP	Earnings-to-price ratio, defined as earnings before extraordinary items (IB) scaled by the market value of equity (PRCC_F × CSHO).
Eret	Excess return, calculated as raw returns minus the one-month Treasury bill rate.
$\lvert FE \rvert$	Analyst forecast accuracy, measured as the average of the mean absolute forecast error of year $t + 1$ earnings sampled at each month of year t.
FF3 Alpha	The intercept estimated from the Fama-French three-factor model regression.
FF4 Alpha	The intercept estimated from the Carhart (1997) four-factor model regression.
FF5 Alpha	The intercept estimated from the Fama-French five-factor model regression.
GMA	Gross profitability, defined as revenues (REVT) minus cost of goods sold (COGS) divided by the lagged total assets.
GrLtNOA	Growth in long-term net operating assets, defined as the annual change in net property, plant, and equipment (PPENT) plus the change in intangibles (INTAN) plus the change in other long-term assets (AO) minus the change in other long-term liabilities (LO) plus the depreciation and amortization expenses (DP), scaled by average total assets.
GrInv	Growth in inventory (INVT).
GrInvest	Growth in capital expenditure (CAPX).
Growth	Average percentage growth in sales (SALE) over year $t - 4$ to year t.
IndTag	Percentage of industry-specific tags used on three financial statements. Based on the FASB taxonomy, we classify a tag as industry-specific if it is related to an Accounting Standards Codification topic in the 900 (Industry) area.
Inv_TA	Inverse of total assets (AT).
Lev	Leverage ratio is defined as total liabilities (LT) divided by the market value of equity (PRCC_F × CSHO).
Log(TA)	The natural logarithm of total assets (AT).
MA	An indicator variable that equals one if the firm is engaged in merger and acquisition during the current year reported by the SDC database, and zero otherwise.
MOM_1m	Cumulative return of month $t - 1$.
MOM_12m	Cumulative return over the 11 months ending one month before month t.
MOM_36m	Cumulative return from month $t - 36$ to month $t - 13$.
MTB	Ratio of market value of equity (PRCC_F × CSHO) to book value of equity (CEQ).
NOA	Net operating assets is computed as operating assets minus operating liabilities. Operating assets are total assets (AT) minus cash and short-term investment (CHE). Operating liabilities are total assets minus debt included in current liabilities (DLC) minus long-term debt (DLTT) minus minority interests (MIB) minus preferred stock (PSTK) minus common equity (CEQ).
NSeg	Number of business segments as reported in Compustat.
NTag	Number of tags used on three financial statements.
OPACC	Operating accruals are measured as − (RECCH + INVCH + APALCH + TXACH + AOLOCH + DPC), where Compustat item RECCH is the decrease (increase) in accounts receivable; INVCH is the decrease (increase) in inventory; APALCH is the increase (decrease) in accounts payable; TXACH is the increase (decrease) in tax payable; AOLOCH is the net change in other current assets; and DPC is the depreciation and amortization from cash flow statement. Researchers have used the first five items (RECCH + INVCH + APALCH + TXACH + AOLOCH) to create a variable approximating the change in operating capital, or OPCAPCH. This accounting concept corresponds to a specific tag, IncreaseDecreaseInOperatingCapital.
Other	Net of all other accruals is calculated as *TACC − (CHG_AR + CHG_INV − CHG_AP − DPExp − AMExp).*
Quick	Quick ratio is defined as the difference between current assets (ACT) and inventory (INVT) divided by current liabilities (LCT).
PPE	Net value of property, plant, and equipment (PPENT), scaled by average total assets.
PROD	Production costs, calculated as COGS plus change in inventory (INVT), scaled by lagged total assets.
RealEstate	Corporate real estate holdings, defined as the sum of buildings (FATB) and capitalized leases (FATL) divided by the gross property, plants, and equipment (PPEGT).
Restructure	An indicator variable for asset restructuring, which equals one if restructuring costs pretax (RCP) are nonzero, and zero otherwise.
RetVol	Standard deviation of monthly returns over the current fiscal year.
⊿Rev	Changes in sales (SALE), scaled by average total assets.
ROA	Return-on-assets, measured as the ratio of income before extraordinary items (IB) to total assets (AT).
Sale	Sales (SALE), scaled by lagged total assets.
SI	Absolute value of special items (SPI), scaled by total assets.
Size	Logarithm of market value of equity (PRCC_F × CSHO) at the end of June.
Suspect_NI	An indicator that equals one if *NI* is greater than or equal to zero but less than 0.005, and zero otherwise, where *NI* is income before extraordinary items (IB) scaled by lagged total assets.
TACC	Total accruals, measured as (IB − OANCF), where Compustat item IB is the income before extraordinary items, and OANCF is cash flows from operating activities.
TB	Taxable income, defined as current tax expenses divided by the maximum federal tax rate, divided by income before extraordinary items (IB). Current tax expenses are measured as the sum of current federal (TXFED) and foreign (TXFO) income taxes. When either of these accounts is missing, current tax expenses are measured as the difference between total income tax expenses (TXT) and the deferred portion of the income tax expenses (TXDI).
$\sigma(EPS)$	Decile ranks of earnings volatility, measured as the standard deviation of EPS over year $t - 4$ to year t, scaled by the share price at the end of year t.

Appendix B. S&P's statements on Compustat

B.1. On Compustat's standardization process

“Standardization is the process of collecting data in a format that removes reporting variability and makes it comparable to other companies. Standardized data is a fundamental necessity when doing company or industry analysis. …

Data is aligned with FASB, SEC, GAAP, etc …, meaning that the models, such as the balance sheet, are in a format that generally is consistent with the accepted forms of financial reporting. It also means that we view the guidance of these entities as being useful in helping with the standardization. An example is with FASB 150, which stipulates that companies with quasi-debt securities that used to be included in the mezzanine section between liabilities and equity, must be broken out between liabilities and/or equity. The amounts that are broken out in liabilities are included in debt and the amounts broken out in equity are kept in the equity section of its models as the FASB has helped guide companies to the correct placement."

Source: Private communication with S&P Global Client Support dated January 22, 2020.

B.2. On the use of XBRL filings by Compustat

"For the collection of North American entities in Compustat, we have not done anything with XBRL as of yet. Fundamentals are still collected using full manual collection. Compustat will not change the way it collects data. Compustat will always look at the data points to verify accuracy.

Compustat's process has not changed due to XBRL. In actuality, the XBRL format has made it more difficult, as Compustat has to convert to HTML to avoid all of the links."

Source: Private communication with S&P Global Client Support dated January 31, 2020.

Appendix C. Details on as-filed data

C.1. Processing as-filed data for matching with Compustat

We retrieve "as-filed" financial statement data from the Financial Statement and Notes Data Sets compiled by the SEC. We also retrieve the annual U.S. GAAP taxonomies for 2009–2019 from the FASB's website.

Each filing references the specific taxonomy (consisting of schema files and relationship files) used in its preparation. The taxonomy defines a hierarchical graph whose nodes are financial statement items and whose arcs identify constituents that may be aggregated (via addition and subtraction) to arrive at the value of that item. We first populate the nodes with values from the filing. Beginning at the bottom of the hierarchy, we then iteratively aggregate values at one level to arrive at values at the next higher level where these are not already provided.

After incorporating custom tags (as described in Appendix C.2 below), our algorithm next considers the following accounting identities:

Assets = AssetsCurrent + AssetsNoncurrent;
Liabilities = LiabilitiesCurrent + LiabilitiesNoncurrent;
CashCashEquivalentsAndShortTermInvestments=CashAndCashEquivalentsAtCarryingValue + ShortTermInvestments;
PropertyPlantAndEquipmentNet = PropertyPlantAndEquipmentGross − AccumulatedDepreciationDepletionAndAmortization PropertyPlantAndEquipment;
IntangibleAssetsNetIncludingGoodwill = IntangibleAssetsNetExcludingGoodwill + GoodWill;
DebtCurrent = ShortTermBorrowings + LongTermDebtAndCapitalLeaseObligationsCurrent.

If any two items in any of these identities are tagged in the filing, and the third is imputed via the bottom-up aggregation, the algorithm replaces the imputed value with the value implied by the accounting identity.

Then, we map Compustat data items to standard taxonomy items by comparing the reporting taxonomy and Compustat's balancing model of financial items. To validate this mapping, we retrieve all firm-year observations from Compustat that have a non-zero value. For each of those observations, we identify the XBRL standard tag whose value is the closest to the Compustat item. We then verify that the most frequently selected tag is indeed the one in the mapping.

Table IA.3 describes the sample selection process. Table IA.4 presents the Compustat-XBRL mapping for the data items examined in this study.

C.2. Treatment of custom tags

When a filer uses tags that are not part of the annually updated GAAP Financial Reporting Taxonomy, the filer defines those tags in an extension taxonomy. Values with such "custom" tags are included in our construction of higher-level financial constructs in four steps.

Step 1: Remove redundant custom tags. For each custom tag, search for its child tag(s) using the firm's self-disclosed extension taxonomy. A custom tag is redundant if all its child tags are standard tags. If one or more child tags are custom tags, then determine whether those custom child tags are redundant. Remove tags iteratively until there are no redundant tags.

Step 2: Determine the parent tag. For each remaining custom tag, find its immediate parent tag according to the extension taxonomy. If the immediate parent tag is a custom tag, then find the parent's parent tag. Stop when the found tag is a standard tag.
Step 3: Create a pool of candidate standard tags. For each non-redundant custom tag, peer tags are the standard tags that are descendants (i.e., child tags, grandchild tags, and so on) of the found parent tag. This is the initial pool of candidate standard tags. Remove from the pool (i) any peer tag (and its descendants) if the filing assigns a value to that tag and (ii) any peer tag having a "debit/credit" balance type different from the custom tag.
Step 4: Identify the nearest equivalent standard tag. For each custom tag, find the best-matching peer tag using keyword matching. Keywords are from the tag's label. Among all resulting custom-peer tag pairs, choose the peer tag with the most common keywords provided there are at least two common keywords and the matching ratio is at least 0.5. The matching ratio is the number of common keywords divided by the lesser of (i) the number of peer tag keywords and (ii) the number of custom tag keywords. In the case of a tie, select the least specific peer tag. If this matching process fails, disregard the custom tag.

C.3. As-filed disclosure quality.

The as-filed DQ score (i.e., $DQ^{As\text{-}filed}$) is constructed in the spirit of Chen et al. (2015); however, it is based on the GAAP reporting taxonomy which is has a more granular hierarchical structure than Compustat's balancing model.[29]

Chen et al. (2015) point out that Compustat reports an item as missing if either the item is applicable but the company does not report it, or the item is inapplicable. Their procedure is intended to capture instances of the former and purge instances of the latter. Analogously, we select XBRL tags corresponding to Chen et al.'s 11 balance sheet and six income statement groups. For each of these grouping tags, we exclude descendant tags with missing values because they are inapplicable and retain all others through a two-step procedure: (i) For each child tag of each group tag, if the child tag and all its descendants have missing values, exclude that child tag and all its descendants; (ii) Exclude any tag (and its descendants) if that tag's peer tags have values aggregated to the value of their (shared) parent tag. The available tags are the remaining descendants of the grouping tag. The number of these tags is denoted *# Available Tags*. The number of these tags with non-missing values is denoted *# Used Tags*.

We construct a DQ score for the balance sheet (i.e., $DQ_{BS}^{As-filed}$) by *value*-weighting 11 balance sheet groups with the following formula:

$$DQ_{BS}^{As-filed} = \sum_{k=1}^{11}\left\{\left(\frac{\#\,Used\ Tags}{\#\,Available\ Tags}\right)_k \times \frac{\$\,Assets_k}{\$\,Total\ Assets}\right\} \div 2,$$

where k denotes the balance sheet group.

Analogously, we classify tags on the income statement into six different groups based on their locations, namely revenue, operating expenses, non-operating income/expenses, interest expense, income tax expense, and income/loss from discontinued operation.[30] Following Chen et al. (2015), the DQ score for the income statement (i.e., $DQ_{IS}^{As-filed}$) by *equal*-weighting six income statement groups:

$$DQ_{IS}^{As-filed} = \sum_{m=1}^{6}\left(\frac{\#\,Used\ Tags}{\#\,Available\ Tags}\right)_m \div 6,$$

where m denotes the income statement group.

After we calculate the as-filed DQ scores for the balance sheet and the income statement separately, we construct a composite DQ score $DQ^{As\text{-}filed}$ by taking the simple average of $DQ_{BS}^{As-filed}$ and $DQ_{IS}^{As-filed}$:

$$DQ^{As-filed} = \left(DQ_{BS}^{As-filed} + \mathrm{DQ}_{IS}^{As-filed}\right) \div 2.$$

Appendix D. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jacceco.2022.101573.

[29] Chen et al. make use of a three-level structure in Compustat's balancing model: 93 balance sheet subaccounts are nested within 25 parent accounts that are nested within 11 group accounts. XBRL tags are much more numerous and are organized into as many as 14 levels. For instance, the XBRL tag corresponding to the Compustat group "current assets" is AssetsCurrent, which is at the first (top) level of the taxonomy. It has 31 s-level child tags. One of these is CashCashEquivalentsAndShortTermInvestments ("cash, cash equivalents, and short-term investments"), which has two third-level child tags: ShortTermInvestments ("short-term investments") and CashAndCashEquivalentsAtCarryingValue ("cash and cash equivalents, at carrying value"). The latter third-level tag has three fourth-level child tags.

[30] Chen et al. (2015) classify income statement items into seven groups based on Compustat's balancing model, which differs from the GAAP reporting taxonomy. As a result, XBRL tags are classified into six groups (instead of the seven groups in Chen et al.).

References

Akbas, F., Markov, S., Subasi, M., Weisbrod, E., 2018. Determinants and consequences of information processing delay: evidence from the Thomson Reuters institutional brokers' estimate system. J. Financ. Econ. 127 (2), 366–388.

Ali, A., Klasa, S., Yeung, E., 2008. The limitations of industry concentration measures constructed with Compustat data: implications for finance research. Rev. Financ. Stud. 22 (10), 3839–3871.

Ball, R., Gerakos, J., Linnainmaa, J., Nikolaev, V., 2016. Accruals, cash flows, and operating profitability in the cross section of stock returns. J. Financ. Econ. 121 (1), 28–45.

Barth, M., Cram, D., Nelson, K., 2001. Accruals and the prediction of future cash flows. Account. Rev. 76 (1), 27–58.

Blankespoor, E., deHaan, E., Marinovic, I., 2020. Disclosure processing costs, investors' information choice, and equity market outcomes: a review. J. Account. Econ. 70 (2–3), 101344.

Bochkay, K., Markov, S., Subasi, M., Weisbrod, E., 2022. The roles of data providers and analysts in the production, dissemination, and pricing of street earnings. J. Account. Res. 60 (5), 1695–1740.

Boritz, J.E., No, W.G., 2020. How significant are the differences in financial data provided by key data sources? A comparison of XBRL, Compustat, Yahoo! Finance, and Google Finance. J. Inf. Syst. 34 (3), 47–75.

Bostwick, E.D., Lamber, S.L., Donelan, J.G., 2016. A wrench in the COGS: an analysis of the differences between cost of goods sold as reported in Compustat and in the financial statements. Account. Horiz. 30 (2), 177–193.

Bushman, R.M., Lerman, A., Zhang, X.F., 2016. The changing landscape of accrual accounting. J. Account. Res. 54 (1), 41–78.

Carhart, M., 1997. On persistence in mutual fund performance. J. Financ. 52 (1), 57–82.

Chen, S., Miao, B., Shevlin, T., 2015. A new measure of disclosure quality: the level of disaggregation of accounting data in annual reports. J. Account. Res. 53 (5), 1017–1054.

Chen, H., Cohen, L., Gurun, U., 2021. Don't take their word for it: the misclassification of bond mutual funds. J. Financ. 76 (4), 1699–1730.

Chychyla, R., Kogan, A., 2015. Using XBRL to conduct a large-scale study of discrepancies between the accounting numbers in Compustat and SEC 10-K filings. J. Inf. Syst. 29 (1), 37–72.

Davis, J., 1994. The cross-section of realized stock returns: the pre-COMPUSTAT evidence. J. Financ. 49 (5), 1579–1593.

De Franco, G., Kothari, S., Verdi, R., 2011. The benefits of financial statement comparability. J. Account. Res. 49 (4), 895–931.

Dechow, P., 1994. Accounting earnings and cash flows as measures of firm performance: the role of accounting accruals. J. Account. Econ. 18, 3–42.

Dechow, P., Dichev, I., 2002. The quality of accruals and earnings: the role of accrual estimation errors. Account. Rev. 77, 35–59.

Dechow, P., Ge, W., Schrand, C., 2010. Understanding earnings quality: a review of the proxies, their determinants and their consequences. J. Account. Econ. 50 (2–3), 344–401.

D'Souza, J.M., Ramesh, K., Shen, M., 2010. The interdependence between institutional ownership and information dissemination by data aggregators. Account. Rev. 85 (1), 159–193.

Erhardt, J., 2016. Remarks at the 2016 AICPA National Conference on current SEC and PCAOB developments. Available at: https://www.sec.gov/news/speech/erhardt-2016-aicpa.html.

Fairfield, P., Whisenant, J., Yohn, T., 2003. Accrued earnings and growth: implications for future earnings performance and market mispricing. Account. Rev. 78 (1), 353–371.

Fama, E., French, K., 1997. Industry costs of equity. J. Financ. Econ. 43 (2), 153–193.

Fama, E., French, K., 2015. A five-factor asset pricing model. J. Financ. Econ. 116, 1–22.

FASB, 2018. Concepts Statement No. 8: Conceptual Framework for Financial Reporting—Chapter 3, Qualitative Characteristics of Useful Financial Information. Amended 08/2018 (Issued 09/2010). FASB, Norwalk, CT.

Green, J., Hand, J., Soliman, M., 2011. Going, going, gone? The apparent demise of the accruals anomaly. Manag. Sci. 57 (5), 797–816.

Green, J., Hand, J., Zhang, X.F., 2017. The characteristics that provide independent information about average US monthly stock returns. Rev. Financ. Stud. 30 (12), 4389–4436.

Hoitash, R., Hoitash, U., 2018. Measuring accounting reporting complexity with XBRL. Account. Rev. 93 (1), 259–287.

Hoitash, R., Hoitash, U., Morris, L., 2021. eXtensible business reporting language: a review and directions for future research. Audit J. Pract. Theor. 40 (2), 107–132.

Hong, H., Lim, T., Stein, J., 2000. Bad news travels slowly: size, analyst coverage, and the profitability of momentum strategies. J. Financ. 55 (1), 265–295.

Hou, K., Xue, C., Zhang, L., 2020. Replicating anomalies. Rev. Financ. Stud. 33 (5), 2019–2133.

Kaplan, Z., Martin, X., Xie, Y., 2021. Truncating optimism. J. Account. Res. 59 (5), 1827–1884.

Kern, B.B., Morris, M.H., 1994. Differences in the Compustat and expanded Value Line databases and the potential impact on empirical research. Account. Rev. 69 (1), 274–284.

Lev, B., Nissim, D., 2004. Taxable income, future earnings, and equity values. Account. Rev. 79 (4), 1039–1074.

Linnainmaa, J., Roberts, M., 2018. The history of the cross-section of stock returns. Rev. Financ. Stud. 31 (7), 2606–2649.

Livnat, J., López-Espinosa, G., 2008. Quarterly accruals or cash flows in portfolio construction? Financ. Anal. J. 64 (3), 67–79.

McLean, R.D., Pontiff, J., 2016. Does academic research destroy stock return predictability? J. Financ. 71 (1), 5–32.

McNichols, M., 2002. Discussion of "The quality of accruals and earnings: the role of accrual estimation errors". Account. Rev. 77, 61–69.

Merrill Corporation, 2016. The SEC's Increasingly Sophisticated Use of XBRL-Tagged Data. Featured interview.

Nallareddy, S., Sethuraman, M., Venkatachalam, M., 2020. Changes in accrual properties and operating environment: implications for cash flow predictability. J. Account. Econ. 69 (2–3), 101313.

Ou, J., Penman, S., 1989. Financial statement analysis and the prediction of stock returns. J. Account. Econ. 11 (4), 295–329.

Payne, J.L., Thomas, W.B., 2003. The implications of using stock-split adjusted I/B/E/S data in empirical research. Account. Rev. 78 (4), 1049–1067.

PricewaterhouseCoopers (PwC), 2014. How Companies Can Minimize Reporting Risks and Realize Benefits—XBRL Submission and Processes.

Richardson, S., Tuna, I., Wysocki, P., 2010. Accounting anomalies and fundamental analysis: a review of recent research advances. J. Account. Econ. 50 (2–3), 410–454.

Roychowdhury, S., 2006. Earnings management through real activities manipulation. J. Account. Econ. 42 (3), 335–370.

Schaub, N., 2018. The role of data providers as information intermediaries. J. Financ. Quant. Anal. 53 (4), 1–34.

Securities and Exchange Commission (SEC), 2009. Interactive data to improve financial reporting. Final rule. Available at: https://www.sec.gov/rules/final/2009/33-9002.pdf.

Sloan, R., 1996. Do stock prices fully reflect information in accruals and cash flows about future earnings? Account. Rev. 71 (3), 289–315.

S&P Global, 2018. The Impact of Disparate Data Standardization on Company Analysis (White paper).

Willis, M., 2013. Who is using XBRL? The XBRL Canada Blog. Available at: http://xbrlca.blogspot.ca/2011/03/who-is-using-xbrlby-mike-willis.html.

Xie, H., 2001. The mispricing of abnormal accruals. Account. Rev. 76 (3), 357–373.