Lost in Standardization:
Revisiting Accounting-based Return Anomalies Using As-filed Financial Statement Data

Kai Du
Smeal College of Business
Penn State University
kxd30@psu.edu

Steven Huddart
Smeal College of Business
Penn State University
huddart@psu.edu

Xin Daniel Jiang
School of Accounting and Finance
University of Waterloo
daniel.jiang@uwaterloo.ca

This draft: September 2021

We thank Changling Chen, James Choi, Lauren Cohen, Ying Compton, Michael Dambra, Kurt Gee, Walter Hamscher, Byoung-Hyoun Hwang, Juhani Linnainmaa, Henock Louis, Julie Marlowe, Jed Neilson, Doron Nissim, Inho Suk, Shuyang Wang, Christine Wiedman, Mike Willis, Biqin Xie, Michael Zhang, and seminar participants at the Securities and Exchange Commission (DERA), Penn State, the University at Buffalo, the University of Waterloo, and XBRL US for helpful comments and suggestions. We are especially grateful to Dan Givoly for many discussions. We acknowledge our respective institutions for financial support.

Lost in Standardization:
Revisiting Accounting-based Return Anomalies Using As-filed Financial Statement Data

Abstract

SEC-mandated, machine-readable structured filings, or “as-filed data,” are an alternative source to Compustat for companies’ accounting data. Discrepancies between as-filed and Compustat data, potentially a result of Compustat’s standardizations, affect inferences about the existence and magnitude of the accruals anomaly: accruals calculated from as-filed data do predict returns and accruals calculated from Compustat data do not. Trades of hedge funds that download structured filings correlate with the as-filed accruals signal and, especially, the discrepancy between as-filed and Compustat accruals signals. Inferences about four other accounting-based anomalies are similarly affected by discrepancies between data sources.

Keywords: asset pricing tests, accruals anomaly, data aggregators, standardization, XBRL

JEL Codes: C81, G14, G23, M41, M48

1. Introduction

Financial statement data assembled by third-party, commercial data aggregators (e.g., Compustat from S&P Global) are the basis for the trading decisions of investment professionals as well as hundreds of published studies on stock return anomalies (e.g., Green, Hand, and Zhang, 2017; Linnainmaa and Roberts, 2018; Hou, Xue, and Zhang, 2020). Data aggregators' efforts to achieve consistency, over time and across companies, in the data they extract from SEC registrants' 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, including a former Deputy Chief Accountant of the Securities and Exchange Commission (SEC) (Erhardt, 2016), who 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 and the data provided by different aggregators.[1] We provide evidence that discrepancies between SEC filings and data aggregators' products are widespread and big enough to affect inference and decision making.

Until recently, it has been infeasible to systematically evaluate the asset-pricing 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-

[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).

mandated structured disclosures in eXtensible Business Reporting Language (XBRL) format, which is 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 asset-pricing implications of data aggregators' standardization practices. We assemble as-filed financial statement data that are analogous to Compustat data items for 19,615 firm-years between 2012 and 2018 and reexamine accounting-based stock return anomalies—with a focus on the accruals anomaly—using these data. We study whether Compustat and as-filed data yield the same inferences about the existence and magnitude of these anomalies.

There are three advantages to using as-filed data, as opposed to using third-party data provided by data aggregators. First, as-filed data are more granular than aggregators' data. For example, Compustat's Fundamental Annual dataset contains about 900 data items. By comparison, according to the 2020 edition of the taxonomy, there are 643 unique balance sheet tags, 574 unique income statement tags, and 766 unique cash flow statement tags.[3] Second, as-filed data adhere to an authoritative, public taxonomy. Because they are not subject to a data aggregator's adjustments, they are verifiable and reproducible. This feature prevents the information loss that occurs when

[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] These counts only pertain to the numerical portion of the financial statements. A typical XBRL filing also contains a large number of disclosure tags that tie text passages to specific disclosure topics (e.g., inventory policies).

data aggregators reduce a filing to a smaller number of data items. Also, they afford users greater flexibility in constructing non-standard measures. For example, modified definitions of GAAP net income (sometimes known as "street" measures) can be readily and systematically calculated from as-filed data. 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. This eliminates the lag between when a corporation files its Form 10-K and when investors can obtain all the quantitative information in the filing in a format suitable for statistical analysis.[4]

To illustrate data sources' impact on asset pricing research and investment practice, we focus on one of the most studied accounting-based stock return regularities, the accruals anomaly (Sloan, 1996). We first document significant discrepancies between the as-filed data and the Compustat data for several accounting items involved in calculating operating accruals. These discrepancies tend to be greater for firms that are smaller or are experiencing higher growth. Furthermore, discrepancies tend to be larger when (i) it is more difficult to compare the accounting practices between industry peers; (ii) the filing contains more industry-specific XBRL tags; or (iii) the financial statements present more granular items (e.g., the change in accounts payable to related parties). In other words, in cases where the accounting is complex or the registrant discloses uncommon accounting items, greater data discrepancies arise from Compustat's standardizations.

We then use the as-filed accruals measure to sort firm-years into portfolios and test whether the low-accruals portfolio has higher returns than the high-accruals portfolio. About 32% of the stocks in a portfolio formed using as-filed data are distinct from stocks in the corresponding

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

portfolio formed using Compustat data. Differences in portfolio composition have the potential to affect portfolio returns and consequent inferences. We find that, when using the as-filed data, the accruals anomaly *is* significant (i.e., the low-accruals portfolio has significantly higher returns than the high-accruals portfolio). When using the Compustat data to compute the accruals measure over the same 2012 to 2018 period, however, *no* accruals anomaly is detected. We confirm these findings using control hedge portfolio analysis and Fama-MacBeth cross-sectional regressions.

We next examine whether institutional investors use as-filed data. If institutional investors are aware of the return-predictive power of as-filed accounting numbers, we expect them to trade in the direction as prescribed by as-filed return-predictive signals rather than by the Compustat returns-predictive signals. Using information acquisition records drawn from the EDGAR Log File Data, we find that hedge funds that acquire as-filed data change their portfolio holdings in a manner consistent with the return-predictive signals of as-filed data, rather than with the analogous Compustat-based signals.

We conduct additional tests to rule out other explanations and to explore whether our findings generalize to other data aggregators and other accounting-based anomalies. First, we examine whether the difference in the documented accruals anomaly across data sources is driven by the fact that Compustat restates accounting items over time. Using *unrestated* Compustat data, we still find a significant difference between the as-filed accruals anomaly and the Compustat-based accruals anomaly.

Although regarded as more faithful than third-party data, as-filed data nevertheless may contain errors and use custom tags that are not defined by the taxonomy. To address the concern that our anomaly findings are driven by data quality issues in as-filed data rather than by Compustat standardizations, we replicate our main analysis (i) after we incorporate custom tags in the calculation of as-filed accruals; and (ii) using a restricted sample that exclude firm-years that contain an accruals-related error in violation of the data quality rules developed by XBRL US, a not-for-profit organization that supports the implementation and adoption of XBRL. Test results for the restricted samples are similar to those for the full sample, suggesting that our results are not driven by data quality issues in as-filed data.

We also repeat our main analysis by substituting Compustat data with FactSet data, another financial statement data vendor with a significant market share. We document the existence of the accruals anomaly based on FactSet data—which contrasts with Compustat—but the magnitude is smaller than that of as-filed data. This result suggests divergent standardization practices among commercial data aggregators.

Finally, our findings lead us to question whether the difference in the documented accruals anomaly also exists among other accounting-based anomalies. Therefore, we revisit 19 other accounting-based anomalies examined in two comprehensive studies, Green et al. (2017) and Hou et al. (2020). We find that the discrepancies between as-filed and Compustat data affect the predictive power of four other accounting-based return predictors: earnings before depreciation and extraordinary items-to-debt ratio (Ou and Penman, 1989), growth in long-term net operating assets (Fairfield, Whisenant, and Yohn, 2003), operating profitability (Fama and French, 2015; Ball et al., 2016), and taxable income (Lev and Nissim, 2004). These four predictors, as well as operating accruals, relative to the remaining 15 predictors, involve data items that are more deeply embedded in the financial statements. In other words, the adjustments made by Compustat seem to be most consequential when Compustat faces greater task complexity in its data collection process.

In accounting and finance, there are many studies on accounting-based anomalies, especially the accruals anomaly. Recent research in this area examines whether anomalies attenuate as more capital is deployed in trading strategies designed to exploit the anomaly (Green, Hand, and Soliman, 2011; McLean and Pontiff, 2016). While prior studies have invariably used financial statement data from Compustat, our study demonstrates that conclusions are sometimes contingent on the data source. In particular, the use of data drawn from structured disclosures may contradict basic propositions on which investors, academics, and preparers of financial statements currently place great importance. Therefore, expanding the use of XBRL and further increasing its reliability (e.g., by providing audit assurance) and usability (e.g., by creating more capable application programmer interfaces or APIs) merits further exploration.

Data aggregators parse regulatory filings and other public information to produce readymade datasets for practitioners and academics. Several studies document associations between data aggregators' dissemination of financial information and investors' reactions to such information (e.g., D'Souza et al., 2010; Schaub, 2018; Akbas et al., 2018). Additional studies scrutinize the integrity or quality of other data aggregation products, including analyst forecast data from I/B/E/S (Kaplan, Martin, and Xie, 2020), and mutual fund performance data from Morningstar (Chen, Cohen, and Gurun, 2020). In the particular 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, Klasa, and Yeung, 2008), and made comparisons with Value Line (Kern and Morris, 1994).[5]

[5] Three earlier studies present account-specific, small-sample, or one-year only evidence of the discrepancies between Compustat and XBRL filings (e.g., Bostwick, 2016; Boritz and No, 2020; Chychyla and Kogan, 2015). None of the studies systematically utilizes the FASB taxonomy to prepare as-filed data.

Ours is the first study to systematically examine how financial statement data gathered using conventional aggregation methods, which rely on human interpretation of complex regulatory filings, compare with data extracted directly from structured filings. Following our methodology, a researcher would be able to assemble as-filed data. That same researcher, however, typically cannot reconcile certain Compustat data entries with the original financial statements, because Compustat's standardization procedures are complex. We show that using structured data in place of third-party data, such as Compustat, affects the practical decision of what stocks to include in an investment portfolio and inferences about the existence of a pricing anomaly. These findings are pertinent to ongoing debates over whether and how to implement data-gathering mandates in other contexts.[6] The availability of structured disclosures calls into question the continuing value of proprietary, and therefore somewhat opaque, standardizations embedded in some data aggregators' products.

The remainder of the paper is organized as follows. Section 2 describes the data and reports descriptive statistics. Section 3 presents the baseline analysis using portfolio analysis and cross-sectional regressions. Section 4 studies hedge funds' trading behavior with respect to as-filed accounting signals. Section 5 conducts additional analysis. Section 6 concludes.

[6] Other regulators have also mandated or are contemplating the implementation of XBRL reporting for their registrants. In 2005, the Federal Financial Institutions Examination Council began requiring all bank institutions under its jurisdiction to provide quarterly Reports of Condition and Income (Call Report) in the XBRL format. The European Central Bank has required national competent authorities to provide XBRL filing indicators when submitting supervisory data points. The Federal Energy Regulatory Commission is adopting a structured data approach to its regulatory reporting system. Also, the Grant Reporting Efficiency and Agreements Transparency Act requires more than 35 federal agencies to modernize their grant reporting systems, possibly by implementing XBRL.

2. Data and Descriptive Statistics

2.1. As-filed financial statement data vs. Compustat

Our "as-filed" financial statement data are based on the Financial Statement and Notes Data Sets compiled by the SEC, which contain financial statement information extracted from periodic corporate XBRL filings, without any aggregation or standardization.[7] For some registrants, data are available from 2009; however, we focus on the period 2012–2018, during which *all* SEC registrants were required to submit periodic filings in XBRL format.[8] When more than one annual filing (10-K or 10-K/A) exists for the same fiscal year, we use the most recent filing before the portfolio formation date for each year.[9]

A major step of our data preparation process involves constructing an as-filed data set that is comparable to the annual fundamental file compiled by Compustat. To do this, we use the authoritative U.S. GAAP Financial Reporting Taxonomy (XBRL) provided by the Financial Accounting Standards Board (FASB).[10] The FASB's reporting taxonomy describes the hierarchical relations among all standard XBRL tags. A high-level tag, or "parent tag," can have many "children;" a child tag can have its own children, and so forth. Parent and child tags are explicitly connected by the calculation links, which are essentially a set of hierarchical relations (e.g., assets include both current assets and non-current assets) and accounting identities (e.g., assets equal the sum of liabilities and stockholders' equity).

[7] See https://www.sec.gov/dera/data/financial-statement-and-notes-data-set.html.

[8] 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).

[9] We examine the potential impact of amended filings and restated financial statements in Section 5.1.

[10] A taxonomy defines tags that identify a datum, its attributes, and its relationships to other data. In XBRL, an associated calculation linkbase organizes monetary elements so that lower-level elements sum up to or are subtracted from one another to yield an upper-level concept. The taxonomy is available at https://www.fasb.org/xbrl.

For each of the Compustat data items that are summed to form accruals (and other return predictors described in Section 5.4), we identify the highest-level tags in the FASB taxonomy that correspond to the Compustat data items based on Compustat' 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 tag (or the combination of several tags) selected dominates all other tags when following a procedure detailed in Appendix C.1.

Occasionally, however, the filing does not contain a value for a high-level tag. In such cases, and consistent with the purpose of the calculation linkbase, we use the hierarchical relations specified by the linkbase to impute the high-level tag value from the values of the appropriate child tags. Appendix C.1 contains further details on this step of our procedure. We emphasize that this step does not involve subjective judgments on our part because the calculation linkbase encodes relationships among all standard tags as determined by the authoritative taxonomy.[11]

In the extant accounting and finance literature (e.g., Hribar and Collins, 2002; Ball et al., 2016), the Compustat measure of operating accruals is typically an aggregate constructed from six data items: – (*recch* + *invch* + *apalch* + *txach* + *aoloch* + *dpc*). Compustat defines item *recch* as the decrease (increase) in accounts receivable, *invch* as the decrease (increase) in inventory, *apalch* as the increase (decrease) in accounts payable, *txach* as the increase (decrease) in tax payable, *aoloch* as the net change in other assets and liabilities, and *dpc* as the depreciation and amortization from cash flow statement. The first five items in this aggregate have long been used by researchers to create a variable approximating the change in operating capital, or Δ*OpCap*. This accounting concept corresponds to a specific tag in the FASB's taxonomy,

[11] Some filings contain non-standard tags. In Section 5.2, we show that our results are not affected by these tags.

IncreaseDecreaseInOperatingCapital, which greatly simplifies and standardizes the extraction of this amount from a set of financial statements filed with the SEC since its structured data mandate came into force in 2009.

Our method does not require us to exactly match the components of operating accruals between Compustat and as-filed data. However, to understand the discrepancy between as-filed and Compustat data, we decompose *IncreaseDecreaseInOperatingCapital* into components that correspond to Compustat data items. Table C.1 presents the details. For example, the Compustat item *invch* is mapped to the tag *IncreaseDecreaseInInventories*. Note that the FASB taxonomy is structured so that *IncreaseDecreaseInOperatingCapital* is atop a hierarchy of child tags, including *IncreaseDecreaseInInventories* among many others. Because FASB's taxonomy does not contain a single high-level tag corresponding to Compustat's *aoloch*, we sum several child tags, using the procedure as described in Appendix C.1, to form the XBRL analog of *aoloch*. All variables are scaled by the total assets (*at*) at the beginning of the fiscal year. Compustat item *at* is mapped to the tag *Assets*; *dpc* is mapped to the tag *DepreciationDepletionAndAmortization*.

Table 1, Panel A reports the number of tags used in the calculation of as-filed operating accruals and its components. For an average firm, 7.483 tags are used to calculate *Accruals*. The numbers of non-missing tags for *recch*, *invch*, *apalch*, *txach*, *aoloch*, and *dpc* are 1.054, 0.647, 1.544, 0.252, 2.560, and 1.426, respectively.[12] Panel B reports the summary statistics for the difference between as-filed and Compustat values for each of the data items used in the calculation of operating accruals. All items are scaled by total assets at the beginning of the fiscal year.[13] The

[12] The mean number of tags used to arrive at Compustat values may be less than 1.0 because not all financial statement items are relevant for all companies. For instance, a company may not have a material inventory balance. In this case, the XBRL tag *IncreaseDecreaseInInventories* (and any child tag of this item) need not appear in the filing.

[13] Total assets are not significantly different between the two data sources.

means of the two corresponding measures are significantly different for most of the components of operating accruals (*Accruals*).[14] Averaging across firm-year observations, Compustat reports a significantly lower mean for *recch* (0.001, t = 3.20), *invch* (0.001, t = 3.76), *apalch* (0.002, t = 4.81), *aoloch* (-0.003, t = -5.26), and *dpc* (-0.005, t = -16.32). *Accruals* is also different across the two data sources (0.005, t = 9.78).

2.2. The nature of Compustat's adjustments

The discrepancy between Compustat data and as-filed financial statements is likely due to Compustat's standardization process, which includes adjustments, aggregations, or 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 remove variation in a same-firm datum over time (S&P Global, 2018).[15] Second, a comparison of the definitions of Compustat's data items with those of the corresponding XBRL tags reveals that Compustat makes a number of 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 read filings and, guided by a proprietary data collection manual, assemble data items by selecting and combining disclosed values from the filings (see Appendix B). This process necessarily involves subjective judgment, which is not a factor when values are computed from tagged structured disclosures using formulas stipulated by the taxonomy. Finally, we note

[14] Compustat defines some changes to conform with the indirect method of computing operating cash flows, meaning that some change variables have the opposite sign to the value obtained by subtracting the lagged value from the current value. We are attentive to and adjust values that, by convention, have opposite signs in Compustat and as-filed data.

[15] "There is a certain amount of 'noise' or variation in data that a highly standardized data source will try to avoid. … In theory, then, a data source with a higher level of standardization should demonstrate trends that show less variance over time. … 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).

that Compustat may misclassify or omit financial statement items during the standardization process, as illustrated in Appendix C.2.

To examine whether company characteristics impact the magnitude of the discrepancy, we sort companies into three groups according to the magnitude of the discrepancy, *Abs*(*Diff_Accruals*), and test whether company characteristics differ across these groups. The characteristics are: firm size (*Size*), measured as the natural logarithm of market capitalization at the end of June of each year; the firm's book-to-market ratio (*BM*); growth in total assets (*AGR*); and cash-based operating profitability (*CbOP*). Detailed definitions of these variables are provided in Appendix A. All variables, unless otherwise specified, are calculated at the fiscal year-end before the portfolio formation date.

Panel C of Table 1 reports the means of firm characteristics for each group. The numbers in each cell are time-series averages of yearly cross-sectional means. Companies for which we observe the largest discrepancies between Compustat and as-filed data tend to have a smaller market capitalization, a higher book-to-market ratio, and lower profitability. We also find that the magnitude of discrepancy is positively correlated with growth, which suggests that Compustat adjustments are greater for high-growth companies, possibly due to the relatively complex nature of their accounting.

The next variable we examine is a measure of accounting comparability (*Comparability*) proposed by De Franco, Kothari, and Verdi (2011), which captures the extent to which two firms produce similar financial statements in similar economic conditions. Compustat's standardization process may lead to larger discrepancies in industries where accounting between firms is less comparable. Consistent with this intuition, Panel C reports that Compustat makes larger adjustments for firms with lower accounting comparability.

Compustat's standardization process may also produce larger discrepancies when the accounting standards applicable to a firm's business are highly industry-specific. We measure industry specificity by the proportion of industry-specific tags (*IndTag*) in the XBRL 10-K filing. Industry-specific tags include tags that are related to Accounting Standards Codification (ASC) Topic Area 900, which provides guidance specific to particular industries (e.g., airlines) or activities (e.g., mining). Panel C shows that discrepancies tend to be larger for companies whose filings contain a larger number of industry-specific tags.

Another step in Compustat's data collection procedure involves a "bottom-up" type of aggregation. Quite often, there are many more data items in the actual 10-K filings than are reported in the Compustat databases. Thus, Compustat has to aggregate these items to arrive at a more standardized menu. To the extent that this "bottom-up" aggregation involves discretion, we suspect that the level of disaggregation or granularity of the original filing is correlated with the magnitude of adjustments made by Compustat.[16] We measure the level of disaggregation by *Depth_CF*, the average "depth" (i.e., the number of parent tags, "grandparent" tags, and so on, that are hierarchically above the given tag in the XBRL taxonomy) of all XBRL tags reported in the statement of cash flows. A greater average depth indicates a greater challenge for the Compustat staff who aggregate the raw items into standardized data items. As reported in Panel C of Table 1, the average depth monotonically increases with the magnitude of the discrepancy. This suggests that Compustat introduces greater adjustments when they need to aggregate more layers of data, a task that could be automated if Compustat relies on XBRL filing and associated taxonomy instead of the conventional document formats (e.g., text or html) as its source of raw data.

[16] Hamscher (2005) emphasized the general principle that XBRL taxonomies should not provide more granularity than the accounting standards they represent.

3. Assessing the Accruals Anomaly Using As-filed Financial Statement Data

Having established that many discrepancies exist between as-filed and Compustat data, that the discrepancies are statistically significant, and that the sizes of the discrepancies are related to company characteristics, we turn to the question of whether these discrepancies are important for asset pricing tests. It might be that the discrepancies are not material enough to affect portfolio formation or inferences on the cross-section of stock returns. However, as we will demonstrate below, this is not the case.

3.1. Portfolio analysis

We first examine whether as-filed 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 accruals 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, Carhart (1997) four-factor alphas, and Fama-French five-factor alphas.

Panel A of Table 2 reports the hedge portfolio returns. For Compustat data, the monthly return to the hedge portfolio that takes a long (short) position in the bottom (top) *Accruals* quintile is not significantly different from zero (using conventional statistical tests) regardless of the return measure, a finding consistent with prior studies that document a gradual attenuation of the accruals anomaly (Green et al., 2011). The excess return to the Compustat hedge portfolio is 0.296% per month. In contrast, the excess return to the hedge portfolio formed based on as-filed accruals is 0.673% per month, more than twice the Compustat raw return. Moreover, the difference in hedge return is significant at the 5% level (0.377%, t = 2.46). Measuring portfolio returns using factor model alphas, we find even greater discrepancies between the two data sources: 0.415%, 0.397%, and 0.416% for the three-, four-, and five-factor models, respectively. The differences in hedge returns are all significant at the 1% level.

To interpret the difference in the hedge returns, we examine (i) the extent to which the two data sets overlap with each other in terms of quintile groupings sorted by *Accruals* (Panel B of Table 2) and (ii) whether the two data sets generate different levels of shuffling from one period to the next, among observations (Panel C of Table 2). Panel B shows that for about 68% of all stock-years, the two data sources place the observations in the same quintile: 15.47%, 12.98%, 12.12%, 12.56%, and 15.09% of all stock-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.

Plainly, it is the observations whose quintile assignment differs by data source (about 32% of all observations) that drive the difference in hedge returns. We control for the overlap in Compustat and as-filed portfolios by constructing a hedge portfolio in which Compustat "disagrees" with as-filed data in the classification of extreme quintiles, in the spirit of control hedge portfolio tests (e.g., Hong, Lim, and Stein, 2000). Specifically, we sort stocks independently based on the two accruals measures. For example, "as-filed Q1" denotes the bottom quintile sorted by the firm's as-filed *Accruals* on June 30 portfolio formation date. We then take a long position in stocks that belong to Compustat Q1 but not as-filed Q1 and a short position in stocks that belong to Compustat Q5 but not 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 Compustat Q1 and a short position in stocks that belong to as-filed Q5 but not Compustat Q5.

Panel D of Table 2 reports the results. When we form portfolios based on the overlap portion of both strategies (i.e., when Compustat agrees with as-filed), we find a positive and marginally significant hedge return measured by *Eret* (0.525%, *t* = 1.85), but an insignificant hedge return measured by factor alphas. When Compustat disagrees with as-filed data, Compustat data generate a negative hedge portfolio return to the accruals strategy (e.g., *Eret*: -0.844%, *t* = -2.72), inconsistent with the accruals anomaly. When as-filed disagrees with Compustat data, as-filed data generate a positive hedge return (e.g., *Eret*: 0.721%, *t* = 1.97). These results confirm that the positive hedge return observed in the as-filed data is largely driven by the disagreement portion of the sample.

3.2. Fama-MacBeth cross-sectional regression analysis

Because hedge portfolio analysis does not accommodate additional controls, 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. These control variables include market risk (*Beta*), firm size (*Size*), the logarithm of book-to-market ratio (*Ln(BM)*), past return momentum over the horizons of one month (*MOM_1m*), 12 months (*MOM_12m*), and 36 months (*MOM_36m*), asset growth rate (*AGR*), and cash-based operating profitability (*CbOP*). Control variables are measured using Compustat data.[17]

Table 3 reports the time-series averages of the monthly cross-sectional regression coefficients and their time-series *t*-statistics. As-filed accruals are negatively associated with future returns (column (1): -2.486, *t* = -2.33). By contrast, Compustat-based accruals are not significantly

[17] Measuring control variables using as-filed data yields qualitatively the same results.

associated with future returns (column (2): 0.306, $t = 0.25$). When both measures of accruals are included in the regression, we find that only as-filed accruals negatively predict stock returns: $Accruals^{Filed}$ is negatively associated with future returns (-3.611, $t = -2.99$), but $Accruals^{Compustat}$ is positively associated with future returns (3.085, $t = 2.46$). These findings are consistent with the portfolio analysis reported in Panel D of Table 2.

Overall, both portfolio analysis and Fama-MacBeth cross-sectional regressions suggest that the accruals anomaly exists when accruals are computed using the as-filed financial statement data but not when accruals are computed using Compustat data.

4. The Use of As-filed Financial Statement Data by Institutional Investors

4.1. Identifying institutional investors that use as-filed financial statement data

Instead of relying on data from commercial vendors, investors may base their trades on data extracted directly from structured financial statements. Large institutions may also develop in-house technologies that automate the collection of financial statement data. Although this trend began even before the advent of structured disclosures, the XBRL mandate facilitated automated access to data necessary to conduct financial analysis. As stated by the SEC in its final rule, "In [XBRL], financial statement information could be downloaded directly into spreadsheets, analyzed in a variety of ways using commercial off-the-shelf software, and used within investment models" (SEC, 2009). This view is shared by the investment community (CFA Institute, 2009). In this section, we examine whether and how institutional investors directly use structured filings to inform their trading decisions.

The EDGAR Log File data allow us to observe when an XBRL filing is accessed. As detailed in Appendix D, we link the IP addresses in the EDGAR Log File Data to institutional investors covered by Thomson Reuters to produce a record of the viewing activities of 871 institutional investors over the period from January 1, 2003 to June 30, 2017. Figure 1 plots the time-series trends in the percentage of institutional investors that access XBRL filings (Panel A) and the total portfolio size under the management of these investors (Panel B). Relatively few institutional investors rely on XBRL filings downloaded directly from EDGAR. The total market value of stock holdings of these investors reported on Form 13F is also small relative to the total holdings of all institutional investors.

We use regression analysis to examine the determinants of institutional investors' viewing of XBRL filings in a given quarter (*Viewing*). We focus on the following characteristics of institutional investors: assets under management (*Log(PortSize)*), age (*Log(Age)*), portfolio turnover (*Turnover*), the concentration of the portfolio (*PortHHI*), portfolio return (*PortRet*), the absolute value of fund flow (|*Flow*|), portfolio return volatility (*PortVol*), and whether the institutional investor has viewed XBRL filings in the prior quarter (*PriorViewing*).

Table 4 reports the regression results. We find that the propensity to view XBRL filings is persistent over time. Large institutions and institutions that perform strongly in the past are more likely to download and view XBRL filings. Furthermore, several characteristics associated with active portfolio management are also associated with the likelihood of downloading and viewing XBRL filings. They include high portfolio turnover and high portfolio concentration as measured by Herfindahl index.

4.2. Institutional investors and the accruals anomaly based on as-filed data

In this section, we study whether institutional investors trade based on as-filed data. We restrict our attention to a sophisticated group of institutional investors, hedge funds. Many hedge funds are active traders who attempt to profit from known anomalies (e.g., Calluzzo, Moneta, and Topaloglu, 2019). Based on untabulated analysis, they are the most active group of users of XBRL filings among all types of institutional investors. We follow Agarwal et al. (2013) to identify registered investment companies that operate hedge funds.[18] Our final sample includes the viewing activities of 247 hedge fund companies.

We test the notion that hedge funds' trades are based on signals from as-filed accruals with the following model:

$$\begin{aligned} TradeValue_{i,j,t+1} &= \delta \cdot Viewing_{i,j,t} + \gamma \cdot Accruals_{j,t} + \beta \cdot Viewing_{i,j,t} \times Accruals_{j,t} + \alpha_i + \omega_j \\ &+ \mu_t + \varepsilon_{i,j,t+1}, \qquad (1) \end{aligned}$$

where *i, j* and *t* index hedge fund company, stock, and quarter (reporting period of institutional investors), respectively. $TradeValue_{i,j,t+1}$ is the change in the market value of the holding position of stock *j* by hedge fund company *i* in quarter *t+1*, inferred from the Thomson 13F holdings data. $Accruals_{j,t}$ is either stock *j*'s most recent as-filed accruals or the difference between stock *j*'s most recent as-filed accruals and Compustat accruals at the end of quarter *t*; α_i, ω_j, and μ_t are hedge fund company, stock, and quarter fixed effects, respectively. The primary variable of interest is $Viewing_{i,j,t} \times Accruals_{j,t}$, whose coefficient β captures the incremental correlation between fund *i*'s trades and the trades suggested by the as-filed accruals signal. If, after viewing as-filed data, hedge funds trade in accordance to the accruals anomaly strategy, the coefficient estimate β should be negative.

We construct a sample of about 120 million hedge fund company-stock-quarter observations. The regression results for testing equation (1) are presented in Table 5. Column (1)

[18] The classification is based on a number of sources, including online business name datasets such as Bloomberg, company websites, and Form ADVs filed by investment companies. Our classification data is based on, but extends that of, Agarwal et al. (2013) to recent years. We thank Agarwal et al. for sharing the data. To avoid double counting, we classify an investment company as a hedge fund company if it manages at least one hedge fund.

shows that the coefficient on as-filed accruals is negative (-0.002, t = -2.16), but the coefficient on *Viewing* × *Accruals* is insignificant, suggesting that on average, downloading XBRL filings does not cause hedge funds to trade in the direction of the accruals strategy. We then decompose as-filed accruals into two components: $Accruals^{Filed} = Accruals^{Compustat} + Diff\_Accruals$. When we include both components and their interactions with *Viewing*, we find that the negative association between *Diff_Accruals* and hedge fund *TradeValue* is more pronounced when the hedge fund company view XBRL filings during this period (-0.547, t = -2.16). Such effect does not exist for the $Accruals^{Compustat}$ component of as-filed accruals.

Taken together, the results of the trading analysis suggest that hedge funds incorporate as-filed accruals information into their trades. Hedge funds seem to be aware of the distinctive return-predictive power of accruals signals embedded in as-filed data. Their trading decisions are aided by direct retrieval of such filings, which may circumvent the shortcomings of commercial data. When as-filed numbers and Compustat numbers disagree, hedge funds seem to rely on the more predictive as-filed numbers.

5. Additional Analysis

5.1. Unrestated Compustat

On occasion, Compustat restates financial statement data after registrants amend their 10-K and 10-Q filings (Livnat and López-Espinosa, 2008). Most academic research is conducted using the regular, standardized version of Compustat data. Our main analysis uses the regular Compusat data, which have been used by the bulk of existing academic research. A small number of other researchers have used either unrestated Compustat or point-in-time Compustat data (e.g., Green et al., 2011). For our analyses, however, it is important to address the possibility that the discrepancies we have identified are caused by these amendments, rather than by Compustat's standardization practices. We thus repeat our baseline analysis by replacing the regular (i.e., "restated") Compustat data with unrestated Compustat data.[19] Once again, on June 30 of each year from 2013 to 2019, we require stocks to have non-missing values for both Compustat and as-filed accruals.

Table 6 reports the results. In Panel A, we find that the difference still exists regardless of the return measure used. For example, using raw excess returns, the difference is 0.283% (t = 1.74). In Panel B, we replicate the Fama-MacBeth cross-sectional regression using unrestated data. We find that Compustat accruals (unrestated) do not predict future returns (column (1): -1.367, t = -0.99). When both unrestated accruals measures are included in the regression, only as-filed accruals negatively predict stock returns (column (2): -2.880, t = -2.41).

5.2. Potential data quality issues with XBRL filings

Like traditional text or html filings, XBRL filings may contain errors or inconsistencies (e.g., Hoitash, Hoitash, and Morris, 2020). To investigate the possibility that the discrepancy between the two data sources is due to errors or inconsistencies in the as-filed data, rather than to Compustat's standardizations, we explore filers' use of custom tags and whether filings violate XBRL US's data quality assertions.

Custom tags. When the standard taxonomy does not accommodate unique circumstances in a filer's particular disclosure, filers are permitted to used custom tags.[20] 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

[19] In general, point-in-time data values are either the unrestated values or the regular Compustat values. Having tested the two cases that lie at the two ends of the spectrum, we can be reasonably assured that using the point-in-time data would yield the same conclusion.

[20] The standard tags are derived from taxonomies available at http://www.sec.gov/info/edgar/edgartaxonomies.shtml.

custom tags.[21] To maintain consistency across companies, we construct as-filed data based only on standard tags from the FASB's taxonomy in our main analysis. To mitigate the concern that the discrepancies may be driven by the exclusion of custom tags, we use the following procedure to identify custom tags that are involved in the calculation of operating accruals.

We first re-produce the Statement of Cash Flows based on the XBRL Taxonomy Extension Presentation Linkbase Document provided by each filer. This document contains one row for each line of the financial statements tagged by the filer. Within the Statement of Cash Flows, we locate the subsection of "Changes in operating assets and liabilities." This subsection usually begins with the textual tag *IncreaseDecreaseInOperatingCapitalAbstract* and ends with the numeric tag *NetCashProvidedByUsedInOperatingActivities*. All custom tags between the two are presumably custom tags related to operating accruals. If we are unable to locate the subsection of "Changes in operating assets and liabilities," we use calculation links provided by filers to identify related custom tags. Specifically, the XBRL Taxonomy Extension Calculation Linkbase Document, which is provided by the filer, contains all calculation relationships among the tags in the filing, including those of custom tags. We use this information to determine whether a custom tag is part of a calculation relationship of operating accruals or a component of operating accruals. Once we have identified the related custom tags, we proceed to construct the alternative version of as-filed data following the same procedure outlined in Appendix C.1.

Table 7, Panel A reports the portfolio analysis using the as-filed data that incorporate custom tags. The hedge returns in the restricted sample are qualitatively similar to those in the

[21] See 17 CFR 232.405(c)(1)(iii)(B): "An electronic filer must create and use a new special element if and only if an appropriate tag does not exist in the standard list of tags for reasons other than or in addition to an inappropriate standard label.". For statistics of recent trends in custom tags, see analysis by the SEC available at https://www.sec.gov/structureddata/gaap_trends_2019.

main analysis in Table 2, Panel A. This result suggests that custom tags (or the exclusion thereof) are unlikely to explain the difference observed in the accruals anomaly.

Errors. We also study a second data quality issue associated with XBRL filings, namely, violations of the data quality rules set forth by XBRL US. Applying these rules to SEC filings identifies the specific tags involved in the violation, as well as other details.[22] We parse these error messages to identify the tags involved in the violation.

Panel B of Table 7 reports the results of the hedge portfolio analysis after excluding filings with errors in accruals-related tags. The hedge returns in the restricted sample are qualitatively similar to those in the main analysis, indicating that errors in XBRL filings do not account for the documented discrepancy in the accruals anomaly. In an untabulated analysis, we exclude firm-years with operating accruals-related custom tags *or* operating accruals-related erroneous tags. The results are also qualitatively the same as those from the main analysis.

In Panel C of Table 7, we replicate the Fama-MacBeth cross-sectional regressions after incorporating custom tags (column (1)), excluding filings with erroneous tags (column (2)), or both (column (3)). The results are qualitatively the same as our main analysis.

5.3. Financial statement data from FactSet

Although financial statement data from Compustat are frequently used in academic research, other data aggregators also offer similar products. To examine whether our inferences generalize to other data aggregators, we replicate the analysis after replacing Compustat data with financial statement data from FactSet, a major competitor of S&P Global.

[22] The data are retrieved from https://xbrl.us/data-quality/filing-results/. We obtain the entire set of violations through an 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 (DQC) ruleset and the xbrlus-cc consistency checks.

The FactSet data items that correspond to *recch*, *invch*, *apalch*, *txach*, *aoloch*, and *dpc* are *ff_receiv_cf*, *ff_inven_cf*, *ff_pay_acct_cf*, *ff_pay_tax_cf*, *ff_wkcap_assets_oth*, and *ff_dep_exp_cf*, respectively. Table S.1 in the Internet Appendix reports the analysis that contrasts FactSet data with as-filed data. Panel A shows that the FactSet counterparts of *recch*, *invch*, *apalch*, *aoloch*, and *dpc* are significantly different from the corresponding as-filed values, even though $Accruals^{FactSet}$ is not different from $Accruals^{Filed}$. Panel B reports the hedge portfolio returns following the accruals strategy. The accruals anomaly is marginally detectable using the FactSet data, although FactSet yields hedge returns that are generally lower and less significant than as-filed data. The Fama-MacBeth cross-sectional analysis, reported in Panel C, shows that in a regression setting, $Accruals^{FactSet}$ does not predict future returns, while $Accruals^{Filed}$ does.

The results using FactSet data indicate that, relative to the accruals measure based on Compustat, the accruals measure based on FactSet seems to deviate less from the as-filed counterpart in terms of the return predictive power. Therefore, data aggregators vary substantially in their standardization practices. This finding highlights the influence of the choice of data source on the inferences regarding the accruals anomaly.

5.4. Other accounting-based anomalies

The accruals anomaly is just one of many accounting-based anomalies documented in prior studies (Richardson, Tuna, and Wysocki, 2010; Green et al., 2017; Hou et al., 2020). We also examine whether our findings generalize to other anomalies. We select all 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 final list of 19 accounting-based return predictors includes: 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 S.2 of the Internet Appendix. For each return predictor, we focus on whether (i) portfolios constructed using Compustat or as-filed data alternatively yield a significant anomaly and (ii) there is a significant difference in the hedge returns between the data sources.

The results for this analysis are summarized in Table 8, Panel A. The hedge portfolio excess returns for predictors *CashDebt*, *GrLtNOA*, *OP*, and *TB* are significantly different between the two data sources. The results of the portfolio analysis for these four predictors are reported in Table S.3 of the Internet Appendix. For *CashDebt, GrLtNOA*, and *TB*, the anomaly finding is stronger using as-filed data than using Compustat data, a pattern also seen in the accruals anomaly.

We then explore the causes of the documented differences in the anomaly findings. We conjecture that the discretionary adjustments made by Compustat are more likely to influence asset pricing tests when the return-predictive signal is based on financial statement items that are more disaggregated and deeper in the financial reporting taxonomy. For accounting items at a more aggregate level (e.g., total assets), Compustat's adjustments are unlikely to play a role. In other words, the discrepancy between the two data sources is more important when examining an "intermediate" accounting item than a "bottom line" item (e.g., net income), or a more aggregate-level item.

To examine this conjecture, for each return-predictive variable, we define the following three variables to capture the nature of the task that Compustat undertakes 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; *Max Depth* is the greatest depth of any tag used to construct the variable. Note that the depth of an income statement (cash flow statement) tag captures the distance from the bottom-line earnings (cash flows), whereas the depth of a balance sheet tag captures the level of granularity of the accounts. A greater depth of the related tags indicates greater complexity in Compustat's standardization process. As reported in Panel B of Table 8, the average *Mean Depth* for return-predictive variables with a significant difference is 5.41, compared to 3.93 for the other variables. The difference, 1.48, is also statistically significant. Similarly, the *Max Depth* for variables with a significant difference is significantly greater than that for other variables. These results support the notion that task complexity drives discrepancies between Compustat'smanual standardization process and as-filed data.

6. Concluding Remarks

Discrepancies between Compustat and as-filed accounting data are large enough to affect inference in asset pricing tests. Using the extensively researched accruals anomaly as a case study, we find that, over a sample period of 2012–2018, portfolios constructed using as-filed data yield significant abnormal returns while portfolios constructed using Compustat data do not. We also find differences for four other accounting-based anomalies. Hedge funds appear to trade on the accruals signals constructed from XBRL data.

Our analysis is conducted in the context of the increasing availability of structured disclosures and the evolving landscape of the data aggregation industry. Underlying these trends are technological advances that make regulatory filings more readily usable by automated algorithms and shorten the distance between preparers and end-users of financial statements. The availability of structured disclosures challenges continued reliance on data aggregators' standardizations, as structured disclosures are, by construction, standardized.[23]

Our study has strong implications for capital markets researchers and investment professionals. For researchers, as-filed financial statement data are different enough from the data produced by Compustat to affect inference in the asset pricing literature. Potentially different findings using as-filed data may warrant new explanations of observed anomalies (or the lack thereof). At the same time, investment professionals can harvest financial data directly from structured disclosures. This may drastically reduce their reliance on data aggregators' standardization and interpretations and thereby reduce data risk.[24]

[23] Our evidence does not, however, imply that as-filed data should replace Compustat in all research applications. Compustat is more readily accessed than as-filed data, a longer time series of Compustat data is available, and Compustat Global provides data for non-U.S. companies that do not adhere to the SEC's structured data standards.

[24] See, for example, "XBRL: A Single Source of Truth," by idaciti, available at https://stories.idaciti.com/choose-your-financial-data-source/.

Appendix A: Definitions of Variables[25]

Variable	Definitions
Accruals	Operating accruals is 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.
Age	The number of years since an institution's first appearance on Thomson Reuters.
AGR	Growth in total assets (*at*).
Beta	Market beta, estimated from a regression of weekly returns on equal-weighted market returns for the previous 3 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 (Compustat item *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 is 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*).
Comparability	Financial statement comparability measure proposed by De Franco et al. (2011). This measure 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 $\hat{\alpha}_i$ and coefficient $\hat{\beta}_i$ on returns ($\hat{\alpha}_j$ and $\hat{\beta}_j$). Then, we calculate the predicted earnings for firm *i* (firm *j*) using $\hat{\alpha}_i$ and $\hat{\beta}_i$ ($\hat{\alpha}_j$ and $\hat{\beta}_j$). We then calculate the comparability for each firm pair (i, *j*), $Comparability_{ijt}$, as the negative value of the average absolute difference between the predicted earnings in the past 16 quarters, divided by 100.

[25] For all accounting-based variables, we only provide a detailed definition for the Compustat-based version. The as-filed version is defined using the same formula as its Compustat-based counterpart, but the data items are constructed from as-filed data. The mappings between Compustat data items and XBRL tags involved in calculating these variables on an as-filed basis are provided in Table C.1 of Appendix C.1 and Table S.2 in the Internet Appendix.

CFP	Cash flows-to-price ratio is defined as cash flows from operating activities (*oancf*) divided by the market value of equity.
Current	Current ratio is defined as current assets (*act*) divided by current liabilities (*lct*).
Depr	Depreciation-to-plant asset ratio is defined as depreciation and amortization expenses (*dp*) divided by net property, plant, and equipment (*ppent*).
Diff_Accruals	Compustat *Accruals* minus as-filed *Accruals*.
Depth_CF	The average depth among standard tags reported in the statement of cash flows. The depth of a tag is the number of layers of "parent tags" above the tag.
dPia	Change in PP&E and inventory-to-assets is defined as the annual change in gross property, plants, and equipment (*ppegt*) plus the change in inventory (*invt*) scaled by 1-year-lagged total assets
EP	Earnings-to-price ratio is defined as earnings before extraordinary items scaled by the market value of equity.
Eret	Excess return, calculated as raw returns minus the one-month Treasury bill rate.
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.
|Flow|	The absolute change in total portfolio value between two consecutive quarters net of the increase dure to returns.
GMA	Gross profitability is defined as revenues (*revt*) minus cost of goods sold (*cogs*) divided by the lagged total assets.
GrLtNOA	Growth in long-term net operating assets is 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*) and plus depreciation and amortization expenses (*dp*), scaled by the average of total assets.
GrInv	Growth in inventory (*invt*).
GrInvest	Growth in capital expenditure (*capx*).
IndTag	The proportion of industry-specific tags in the XBRL 10-K filing. 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.
Lev	Leverage ratio is defined as total liabilities (*lt*) divided by the market value of equity.
MOM_1m	The cumulative return of month *t*–1.
MOM_12m	The cumulative return over the 11 months ending one month before month *t*.
MOM_36m	The cumulative return from month *t*–36 to month *t*–13.
NOA	Net operating assets is computed as operating assets minus operating liabilities. Operating assets are total assets 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*), and minus common equity (*ceq*).
OP	Operating profitability is defined as revenue (*revt*), minus cost of goods sold (*cogs*), minus SG&A expenses (*xsga*) and add R&D expenses (*xrd*), scaled by average total assets.
$\Delta OpCap$	Change in operating capital, calculated as *recch* + *invch* + *apalch* + *txach* + *aoloch*. See the definition of *Accruals* for more details on these Compustat data items.
PortHHI	The Herfindahl index of the portfolio is defined based on the market value of each component stock.
PortRet	The monthly average return on the portfolio during the quarter.
PortSize	Total equity portfolio size is defined as the market value of an institution's quarter-end holdings.
PortVol	The monthly portfolio return volatility during the past 12 months ending in the current quarter-end.
PriorViewing	An indicator which equals 1 if an institutional investor views (i.e., downloads) any 10-K XBRL filings in the previous quarter.
Quick	Quick ratio is defined as the difference between current assets (*act*) and inventory (*invt*) divided by current liabilities (*lct*).
RealEstate	Corporate real estate holdings are defined as the sum of buildings (*fatb*) and capitalized leases (*fatl*) divided by the gross property, plants, and equipment (*ppegt*).
Size	The logarithm of market value of equity at the end of June.
# Tags	The number of tags used to construct the return-predictive variable.
TB	Taxable income is defined as the current tax expenses divided by 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 is measured as the difference between total income tax expenses (*txt*) and the deferred portion of the income tax expenses (*txdi*).
TradeValue	The difference between the market value of stock holding at the end of current quarter and the market value of stock holding at the end of last quarter. Stock holdings are measured at the investment company level, based on Thomson 13F (s34) data.
Turnover	Portfolio turnover is defined as the average quarterly portfolio turnover rate during the past four quarters (ending in the current quarter), or as many quarters as there are available. For an individual quarter, the portfolio turnover rate is computed as the lesser of purchases and sales, divided by the average portfolio size during the quarter.
Viewing	An indicator which equals 1 if an institutional investor views (i.e., downloads) the most recent 10-K XBRL filing before the end of the quarter from EDGAR within a calendar quarter, and the viewing activity is not classified as robot-generated based on a procedure detailed in Appendix D, and 0 otherwise.

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. Although the FASB and SEC regulate reporting practices, there is enough latitude provided to make comparing data difficult from one period to another and from one company to another. An extensively trained staff of industry-specialized experts scours company and SEC reports to give you data that is standardized across time and industries. This eliminates biased reporting methods that distort data.

When the Compustat collection teams standardize data for both U.S. and non-U.S. companies, the researcher goes through the financial statements, notes, management discussion, and other parts of the financial report to extract the data and input it into the various balancing models for the balance sheet, income statement, cash flow statement, and supplemental items that models contain. The process of standardization makes companies as comparable as possible across all industries and countries. An example is that for a specific item, such as trade accounts receivable, there is definition that indicates all of the types of receivables that companies report that are included in that item. If an item is reported that does not fit into that definition, it fits into another item, such as other receivables. Data Researchers use the definitions for each item to determine the proper placement of the data that is reported by each company so that the values for Compustat data are consistent across the board. Companies sometimes have items that involve judgments as to where the best location should be in the Compustat models and this is where the experience of the Data Researchers comes into play as they collaborate and determine the best location based on the description and nature of the item.

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: Matching As-filed Data with Compustat and Sample Selection

C.1. Matching as-filed data 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 taxonomy for the years 2009 to 2019 from the FASB's website.

After determining the version of reporting taxonomy a company has used to prepare its as-filed 10-K, we first document all recorded values for their corresponding tags. We next impute values for the remaining tags. The imputation process starts from tags at the most granular level. At this level, we assign zeros to tags that are part of the taxonomy, but not used by the filer, and add up values of tags that belong to the same parent tag. If the company does not report a value for the parent tag, we assign the calculated value to that tag. Next, we move one level up and focus on the tags at the higher levels, and repeat the above process until we reach tags at the most aggregate level. The imputation process ensures that there is a value for every standard tag. This is an intended use of the taxonomy and the associated calculation linkbase.

Next, we create a mapping between Compustat data items and XBRL standard tags by comparing the reporting taxonomy and Compustat's balancing model of financial items (S&P Global, 2017). We validate the mapping by verifying that the tag (or the combination of several tags) selected dominates all other tags in the following sense. For each Compustat item, we retrieve all firm-year observations from Compustat that have a non-zero value. Then, for each of those observations, we identify the XBRL standard tag whose value is the closest to the Compustat item. We then sample all the selected tags from the observations, and verify that the most frequently selected tag is indeed the one in the mapping we created based on accounting concepts.

Compustat items *recch*, *invch*, *apalch*, and *txach* are mapped to *IncreaseDecreaseInReceivables*, *IncreaseDecreaseInInventories*, *IncreaseDecreaseInAccountsPayableAndAccruedLiabilities*, and *IncreaseDecreaseInAccruedTaxesPayable,* respectively. The Compustat item *aoloch* is an exception because this item is a broad collection of miscellaneous changes in assets and liabilities. We follow the procedure described above to sum over all tags (and their child tags) that are components of *IncreaseDecreaseInOperatingCapital*, but not components of the four aforementioned tags.

Table C.1 presents the Compustat-XBRL mapping for data items used in the calculation of accruals, as well as each component of the change in operating capital. Table C.2 describes the sample selection process.

Table C.1: Matched XBRL Tags

Data Item	Description	XBRL Tag
at	Assets – Total	Assets
Δ*OpCap (=recch + invch + apalch + txach + aoloch)*	Change in Operating Capital	IncreaseDecreaseInOperatingCapital
recch	Accounts Receivable – Decrease (Increase)	IncreaseDecreaseInReceivables
invch	Inventory – Decrease (Increase)	IncreaseDecreaseInInventories
apalch	Accounts Payable and Accrued Liabilities – Increase (Decrease)	IncreaseDecreaseInAccountsPayableAndAccruedLiabilities
txach	Income Taxes – Accrued – Increase (Decrease)	IncreaseDecreaseInAccruedTaxesPayable
aoloch	Assets and Liabilities – Other – Net Change	
dpc	Depreciation and Amortization	DepreciationDepletionAndAmortization

Table C.2. Sample Selection Process

Step	Requirement	Obs.
1	Observations in Compustat with fiscal year end date between 2012 and 2018	48,894
2	Stock returns from CRSP are available	33,490
3	Matching XBRL data with Compustat	28,362
4	Requiring Industry Classification Code (SIC) non-missing and dropping observations from Finance Industry (SIC: 6000–6999) and Utility Industry (SIC: 4900–4999)	21,727
5	Restricting sample to firms listed on NYSE/AMEX/NASDAQ	20,176
6	Requiring operating accruals non-missing from Compustat and XBRL	19,615

C.2. Illustrations of discrepancies between Compustat and as-filed data

Example: DSP Group Inc.

CIK: 0000915778
GVKEY: 029722
Filing Date: March 11, 2020 (FY 2019)
10-K Filing Accession Number: 0001437749-20-004811

"Changes in operating assets and liabilities" from the Consolidated Statements of Cash flows (in thousands):

		For the Years Ended December 31, 2019
Changes in operating assets and liabilities:	**XBRL Tag**	
Deferred income tax assets and liabilities, net	IncreaseDecreaseInDeferredIncomeTaxes	(2,833)
Trade receivables, net	IncreaseDecreaseInAccountsReceivable	(1,916)
Other accounts receivable and prepaid expenses	IncreaseDecreaseInOtherAccountsReceivableAndPrepaidExpenses	817
Inventories	IncreaseDecreaseInInventories	2,345
Long-term prepaid expenses and lease deposits	IncreaseDecreaseInLongTermPrepaidExpensesAndLeaseDeposits	(28)
Trade payables	IncreaseDecreaseInAccountsPayableTrade	(1,197)
Accrued compensation and benefits	IncreaseDecreaseInAccruedCompensationAndBenefits	2,515
Income tax accruals	IncreaseDecreaseInAccruedIncomeTaxesPayable	1,653
Accrued expenses and other accounts payable	IncreaseDecreaseInOtherAccountsPayableAndAccruedLiabilities	(573)
Accrued severance pay, net	IncreaseDecreaseInAccruedSeverancePayNet	84
Accrued pensions	IncreaseDecreaseInPensionPlanObligations	26

Table C.3: Components of Operating Accruals for DSP Group Inc. (FY2019): Compustat vs. As-filed Data

Panel A. Compustat vs. As-filed Values (in millions)

Data Source	*recch*	*invch*	*apalch*	*txach*	*aoloch*
Compustat	0.000	2.345	0.000	1.653	-0.272
As-filed	-1.916	2.345	-1.744	1.653	0.555

Panel B. Analysis of Discrepancies

Financial Statement Line Item		Compustat Data	As-filed Data	XBRL Tag
Trade receivables, net		-	(1.916)	IncreaseDecreaseInAccountsReceivable
	recch	Missing	(1.916)	IncreaseDecreaseInReceivables
Inventories	*invch*	2.345	2.345	IncreaseDecreaseInInventories
Trade payables		-	(1.197)	IncreaseDecreaseInAccountsPayableTrade
Accrued pensions		-	0.026	IncreaseDecreaseInPensionPlanObligations
Accrued expenses and other accounts payable		-	(0.573)	IncreaseDecreaseInOtherAccountsPayableAndAccruedLiabilities
	apalch	Missing	(1.744)	IncreaseDecreaseInAccountsPayableAndAccruedLiabilites
Income tax accruals	*txach*	1.653	1.653	IncreaseDecreaseInAccruedIncomeTaxesPayable
Trade receivables, net		(1.916)		
Trade payables		(1.197)		
Accrued pensions		0.026		
Accrued expenses and other accounts payable		(0.573)		
Other accounts receivable and prepaid expenses		0.817	0.817	IncreaseDecreaseInOtherAccountsReceivableAndPrepaidExpenses
Long-term prepaid expenses and lease deposits		(0.028)	(0.028)	IncreaseDecreaseInLongTermPrepaidExpensesAndLeaseDeposits
Accrued compensation and benefits		2.515	2.515	IncreaseDecreaseInAccruedCompensationAndBenefits
Accrued severance pay, net		0.084	0.084	IncreaseDecreaseInAccruedSeverancePayNet
Deferred income tax assets and liabilities, net		**Omitted**	(2.833)	IncreaseDecreaseInDeferredIncomeTaxes
	aoloch	(0.272)	0.555	
Change in operating capital		3.726	0.893	IncreaseDecreaseInOperatingCaptial
Depreciation and amortization		3.281	3.281	DepreciationDepletionAndAmortization
Operating accruals		(7.007)	(4.174)	

Table C.3 presents the calculation of operating accruals using the two data sources, for DSP Group Inc. Panel A reports the values of the related Compustat items, and their as-filed counterpart. Panel B presents the financial statement items or tags that are input to these values. For *recch*, Compustat omits "Trade receivables, net" and thus reports a missing value. For *apalch*, Compustats omits "Trade payables," "Accrued pensions," and "Accrued expenses and other accounts payable," and reports a missing value. For *aoloch*, Compustat (1) omits "Deferred income tax assets and liabilities, net" and (2) includes the aforementioned four accounts that are elements of *IncreaseDecreaseInReceivables* or *IncreaseDecreaseInAccountsPayableAndAccruedLiabilites*. Some of the classification discrepancies (e.g., Trade receivables, net) only affect the individual components, but not the overall operating accruals. For *invch* and *txach*, both data sources yield the same value. In the table, "-" denotes the case in which Compustat classifies the financial statement line item in the calculation of some other data item (in this example, *aoloch*).

Figure C.1: XBRL Tags Used in Constructing As-filed Data: DSP Group Inc.

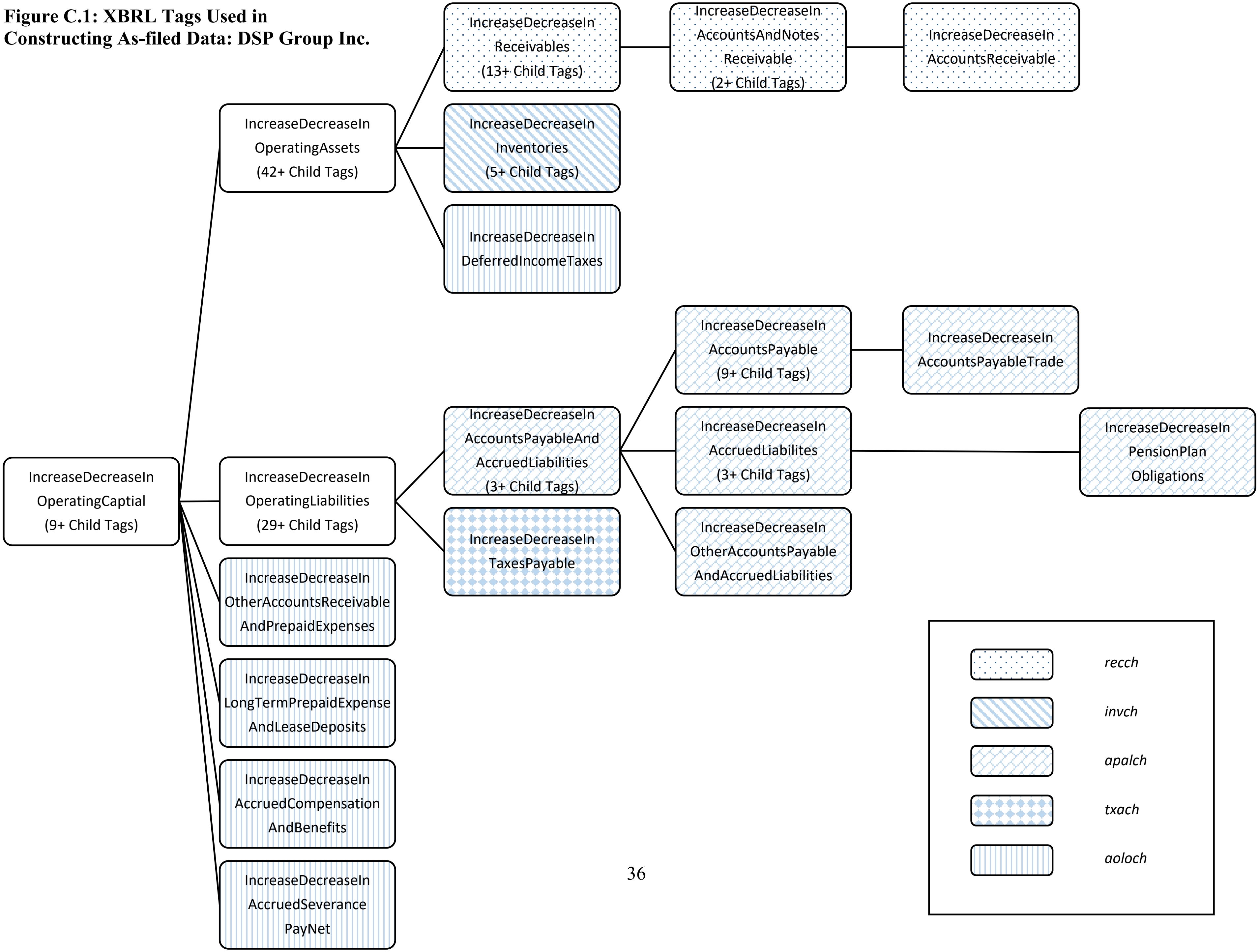

IncreaseDecreaseIn OperatingCaptial (9+ Child Tags)
IncreaseDecreaseIn OperatingAssets (42+ Child Tags)
IncreaseDecreaseIn Receivables (13+ Child Tags)
IncreaseDecreaseIn AccountsAndNotes Receivable (2+ Child Tags)
IncreaseDecreaseIn AccountsReceivable
IncreaseDecreaseIn Inventories (5+ Child Tags)
IncreaseDecreaseIn DeferredIncomeTaxes
IncreaseDecreaseIn OperatingLiabilities (29+ Child Tags)
IncreaseDecreaseIn AccountsPayableAnd AccruedLiabilities (3+ Child Tags)
IncreaseDecreaseIn AccountsPayable (9+ Child Tags)
IncreaseDecreaseIn AccountsPayableTrade
IncreaseDecreaseIn AccruedLiabilites (3+ Child Tags)
IncreaseDecreaseIn PensionPlan Obligations
IncreaseDecreaseIn OtherAccountsPayable AndAccruedLiabilities
IncreaseDecreaseIn TaxesPayable
IncreaseDecreaseIn OtherAccountsReceivable AndPrepaidExpenses
IncreaseDecreaseIn LongTermPrepaidExpense AndLeaseDeposits
IncreaseDecreaseIn AccruedCompensation AndBenefits
IncreaseDecreaseIn AccruedSeverance PayNet
recch
invch
apalch
txach
aoloch

Figure C.1 illustrates the hierarchy of the XBRL tags involved in calculating operating accruals for DSP Group, Inc. To facilitate a comparison with Compustat, we group tags used by the filer that combine to provide the XBRL counterpart to each Compustat item. The grouping is indicated by fill pattern and is indicated in the legend. For example, tags *IncreaseDecreaseInAccountsReceivable*, *IncreaseDecreaseInAccountsAndNotesReceivable*, and *IncreaseDecreaseInReceivables* are related to the calculation of Compustat variable *recch*. In our example, filer reports the value of $1.916 million for the tag *IncreaseDecreaseInAccountsReceivable*, which is the "grand-child tag" of the tag *IncreaseDecreaseInReceivables*. According to imputing procedures elaborated in Appendix C.1, the imputed value for the tag *IncreaseDecreaseInReceivables* is $1.916 million. According to the matching table in Table C.1, the Compustat item *recch* is matched to the XBRL tag *IncreaseDecreaseInReceivables*. As a result, *recch* equals to $1.916 million using as-filed data. Tags at the end of each branch are the ones that appear in DSP Group Inc.'s XBRL filing. However, we also provide the number of tags in the FASB taxonomy that are children of the high-level tag. For example, there are more than 42 child tags of *IncreaseDecreaseInOperatingAssets*. These child tags include 42 standard tags listed in the FASB taxonomy and may include custom tags that are not listed in the taxonomy.

Appendix D: EDGAR Viewing Activities by Institutional Investors

We obtain records of the retrieval of 10-K/10-Q filings from the EDGAR Log File data, which cover the period between January 1, 2003 and June 30, 2017.[26] Each record from the EDGAR Log File data contains the IP address of the requesting user with the fourth octet obfuscated. It also includes the timestamp of the request and the accession number of the filing requested. Institutional investors may access forms 10-K via channels other than EDGAR, such as the filers' websites or through a data vendor. Thus, the number of downloads from EDGAR likely understates the actual number of cases in which institutions access 10-K filings.

We exclude unsuccessful requests and requests that land on index pages. We merge the Log File data with EDGAR index files by accession number to gather information on the form type, filing date and time, and name of the filing entity.

We match the organizations associated with the IP addresses to institutional investors covered by the Thomson Reuters 13F database. Information on organizational IP addresses comes from the Whois database of the American Registry for Internet Numbers (ARIN). We follow Chen et al. (2020) to decipher the fourth octet, an obfuscated IPv4 address from the EDGAR Log File. The matching results in a mapping file between IP addresses and *mgrno* (Thomson's identifier of investment managers).

We further require that viewing activity take place within the quarter after the reporting period for which the 10-K was filed. This requirement ensures that the information contained in the filing is contemporaneous with and relevant to the trading decision of the viewing company.

We also remove robot-generated viewing activities. We classify requests as robot-generated if they are associated with self-identified web crawlers or with daily IP addresses that searched more than 50 unique firms' filings, a criterion also used by Lee, Ma, and Wang (2015).

[26] Available at https://www.sec.gov/dera/data/edgar-log-file-data-set.html.

References

Akbas, F., S. Markov, M. Subasi, and E. Weisbrod. 2018. Determinants and consequences of information processing delay: evidence from the thomson reuters institutional brokers' estimate system. *Journal of Financial Economics* 127 (2), 366–388.

Agarwal, V., W. Jiang, Y. Tang, and B. Yang. 2013. Uncovering hedge fund skill from the portfolio holdings they hide. *Journal of Finance* 68 (2), 739–783.

Ali, A., S. Klasa, and E. Yeung. 2008. The limitations of industry concentration measures constructed with Compustat data: implications for finance research. *Review of Financial Studies* 22 (10), 3839–3871.

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

Bostwick, E. D., S. L. Lamber, and J. G. Donelan. 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. *Accounting Horizons* 30 (2), 177–193.

Boritz, J., and W. No. 2020. How significant are the differences in financial data provided by key data sources? a comparison of XBRL, Compustat, Yahoo! Finance, and Google Finance. *Journal of Information Systems* 34 (3), 47 – 75.

Calluzzo, P., F. Moneta, and S. Topaloglu. 2019. When anomalies are publicized broadly, do institutions trade accordingly? *Management Science* 65 (10), 4555–4574.

Carhart, M. 1997. On persistence in mutual fund performance. *Journal of Finance* 52 (1), 57–82.

CFA Institute. 2009. eXtensible Business Reporting Language: A Guide for Investors. CFA Institute Centre for Financial Market Integrity. Available at https://www.cfainstitute.org/-/media/documents/article/position-paper/xtensible-business-reporting-language-guide-for-investors.ashx.

Chen, H., L. Cohen, and U. Gurun. 2020. Don't take their word for it: the misclassification of bond mutual funds. Working paper.

Chen, H., L. Cohen, U. Gurun, D. Lou, and C. Malloy. 2020. IQ from IP: simplifying search in portfolio choice. *Journal of Financial Economics* 138 (1), 118–137.

Chychyla, R., and A. Kogan. 2015. Using XBRL to Conduct a large-scale study of discrepancies between the accounting numbers in Compustat and SEC 10-K filings. *Journal of Information Systems* 29 (1), 37–72.

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

De Franco, G., S. Kothari, and R. Verdi. 2011. The benefits of financial statement comparability. *Journal of Accounting Research* 49 (4): 895–931.

D'Souza, J. M., Ramesh, K., and Shen, M. 2010. The interdependence between institutional ownership and information dissemination by data aggregators. *The Accounting Review* 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., J. Whisenant, and T. Yohn. 2003. Accrued earnings and growth: Implications for future earnings performance and market mispricing. *The Accounting Review* 78, 353–71.

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

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

Green, J., J. Hand, and X. Zhang. 2017. The characteristics that provide independent information about average us monthly stock returns. *Review of Financial Studies* 30 (12), 4389–4436.

Hamscher, W. 2005. Financial Reporting Taxonomies Architecture 1.0. Recommendation dated 2005-04-25. Available at: http://www.xbrl.org/technical/guidance/FRTA-RECOMMENDATION-2005-04-25.htm.

Hoitash, R., U. Hoitash, and L. Morris. 2020. eXtensible Business Reporting Language: a review and directions for future research. Working paper.

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

Hou, K., C. Xue, and L. Zhang. 2020. Replicating anomalies. *Review of Financial Studies* 33 (5), 2019–2133.

Hribar, P. and D. Collins. 2002. Errors in estimating accruals: implications for empirical research. *Journal of Accounting Research* 40 (1), 105–134.

Kaplan, Z., X. Martin, and Y. Xie. 2020. Truncating optimism. *Journal of Accounting Research*, forthcoming.

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

Lee, C., P. Ma, and C. Wang. 2015. Search-based peer firms: Aggregating investor perceptions through internet co-searches. *Journal of Financial Economics* 116 (2), 410–431.

Lev, B., and D. Nissim. 2004. Taxable income, future earnings, and equity values. *The Accounting Review* 79, 1039–1074.

Linnainmaa, J., and M. Roberts. 2018. The history of the cross-section of stock returns. *Review of Financial Studies* 31(7), 2606–2649.

Livnat, J., and G. López-Espinosa. 2008. Quarterly accruals or cash flows in portfolio construction?. *Financial Analysts Journal* 64 (3), 67–79.

McLean, R.D., and J. Pontiff. 2016. Does academic research destroy stock return predictability? *Journal of Finance* 71 (1), 5–32.

Merrill Corporation. 2016. The SEC's increasingly sophisticated use of XBRL-tagged data. Featured interview.

Ou, J., and S. Penman. 1989. Financial statement analysis and the prediction of stock returns. *Journal of Accounting and Economics* 11, 295–329.

PricewaterhouseCoopers (PwC). 2014. How companies can minimize reporting risks and realize benefits—XBRL submission and processes.

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

Schaub, N. 2018. The role of data providers as information intermediaries. *Journal of Financial and Quantitative Analysis* 53, 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? *The Accounting Review* 71 (3), 289–315.

S&P Global, 2017. Compustat Balancing Model. Available at: https://wrds-www.wharton.upenn.edu/documents/840/Balancing_Model_-_North America Company Data X1BZb7h.xls.

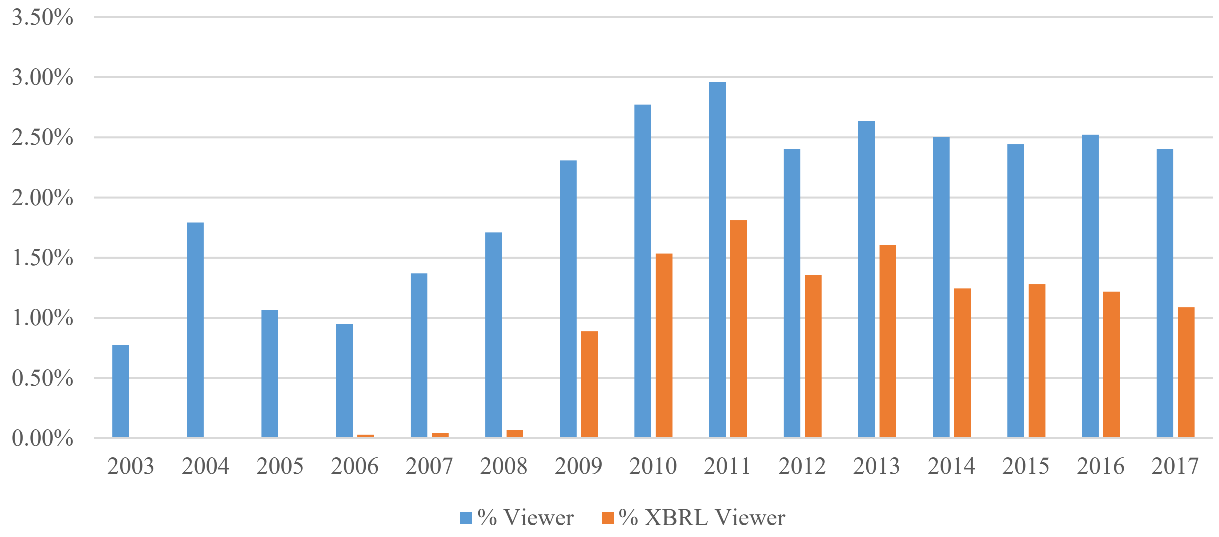
S&P Global, 2018. The impact of disparate data standardization on company analysis. White paper.

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.

Figure 1: Trends in Viewing Activities for Institutional Investors

This figure presents the time-series trends in institutional investors' use of XBRL filings (as-filed financial statement data). In Panel A, "%Viewer" is the percentage of institutional investors that view (i.e., download) any EDGAR filings; "%XBRL Viewer" is the percentage of institutional investors that view any XBRL filings. In Panel B, "Total PortSize" is the total stock holdings of all institutional investors reported on Form 13F; "Viewer PortSize" ("XBRL Viewer PortSize") is the total stock holdings of institutional investors that view (XBRL) EDGAR filings.

Panel A: Percentage of Institutional Investors



Panel B: Total Portfolio Size Under Management (in $ trillions)

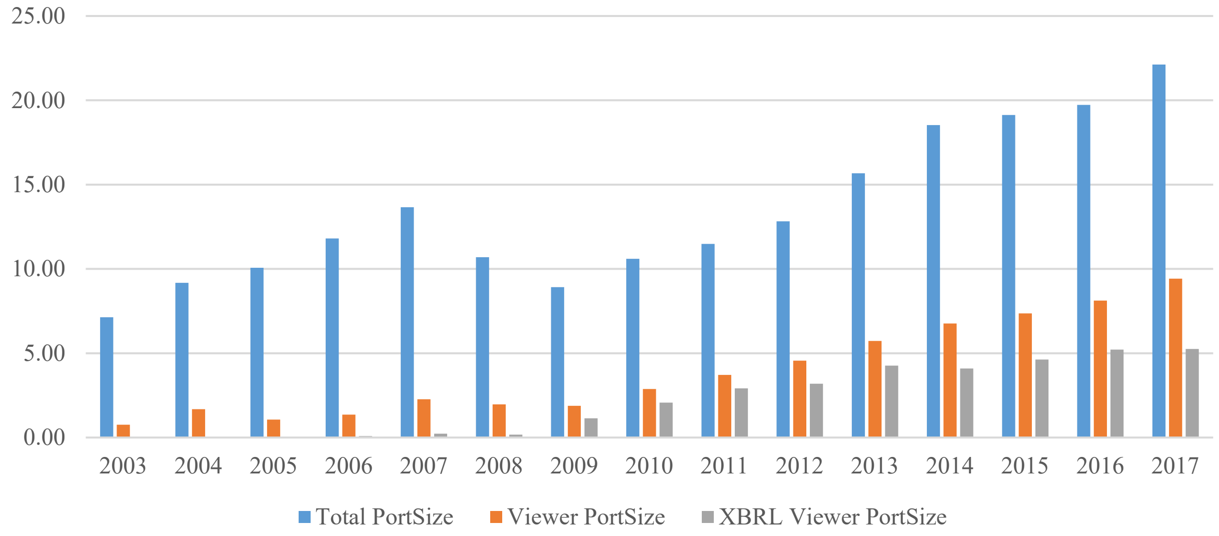


Table 1: Descriptive Statistics

Panel A reports the descriptive statistics on the number of XBRL tags used in calculating as-filed operating accruals as well as their components. Panel B reports the descriptive statistics on the discrepancy between the Compustat and as-filed values for each component in Hribar and Collins' (2002) formula for operating accruals computed over 19,615 firm-years. All variables are scaled by the total assets at the beginning of fiscal year. The standard errors of the two-sample *t*-tests adjust for clustering at the firm and year levels. Panel C reports mean firm characteristics for each accruals adjustment group. On June 30 of each year from 2013 to 2019, we first sort stocks into deciles based on the absolute value of the difference between Compustat-based accruals and as-filed accruals (i.e., |*Diff_Accruals*|). We then categorize stocks into three groups, the bottom 30 percent (*Small* Group), middle 40 percent (*Medium* Group), and top 30 percent (*Large* Group). All firm characteristics, unless otherwise specified, are calculated using Compustat data for the fiscal year-end before the June 30 portfolio formation date. *Size* is firm size measured as the natural logarithm of market capitalization at the end of June of each year. *BM* is the firm's book-to-market ratio. *AGR* is growth in total assets. *CbOP* represents cash-based operating profitability. *Comparability* represents the financial statement comparability. *IndTag* is the percentage of industry-specific tags used in the XBRL filing. *Depth_CF* measures the average level of depth among standard tags reported in the statement of cash flows. The numbers in each cell are time-series averages of yearly cross-sectional means. Significance at the 10%, 5%, and 1% levels (two-sided) are denoted by *, **, and ***, respectively.

Panel A: Number of Tags Used in Calculating As-Filed Operating Accruals

	N	*Mean*	*Std. Dev.*	*Min.*	*Q1*	*Median*	*Q3*	*Max.*
recch	19,615	1.054	0.577	0	1	1	1	7
invch	19,615	0.647	0.478	0	0	1	1	2
apalch	19,615	1.544	0.814	0	1	1	2	5
txach	19,615	0.252	0.445	0	0	0	0	3
aoloch	19,615	2.560	1.319	0	2	2	3	13
dpc	19,615	1.426	0.919	0	1	1	2	7
Accruals	19,615	7.483	2.175	1	6	7	9	18

Panel B: Financial Statement Items Used in Calculating Accruals

	Accruals	*recch*	*invch*	*apalch*	*txach*	*aoloch*	*dpc*
Compustat Data							
Mean	-0.040	-0.010	-0.006	0.008	0.000	0.003	0.046
Std. Dev.	0.075	0.041	0.028	0.037	0.004	0.054	0.038
Q1	-0.071	-0.018	-0.008	-0.000	0.000	-0.011	0.022
Median	-0.035	-0.003	0.000	0.000	0.000	-0.000	0.038
Q3	-0.005	0.003	0.000	0.011	0.000	0.012	0.059
As-filed Data							
Mean	-0.034	-0.010	-0.006	0.009	0.000	-0.001	0.040
Std. Dev.	0.073	0.039	0.025	0.041	0.004	0.044	0.034
Q1	-0.066	-0.017	-0.008	-0.006	0.000	-0.009	0.016
Median	-0.032	-0.002	0.000	0.003	0.000	-0.000	0.034
Q3	-0.002	0.003	0.000	0.017	0.000	0.007	0.054
Two-sample Tests (As-filed Data – Compustat Data)							
Differences in mean	0.005***	0.001***	0.001***	0.002***	0.000	-0.003***	-0.005***
Clustered *t*-stat.	(9.78)	(3.20)	(3.76)	(4.81)	(0.41)	(-5.26)	(-16.32)

Panel C: Firm Characteristics by Magnitude of Discrepancy

|Diff_Accruals|	*N*	*Size*	*BM*	*AGR*	*CbOP*	*Comparability*	*IndTag*	*Depth_CF*
Small	5,881	6.476	0.591	0.117	0.097	-2.602	0.075	6.361
Medium	7,852	6.641	0.628	0.130	0.092	-2.801	0.077	6.393
Large	5,882	5.966	0.611	0.346	0.038	-3.378	0.080	6.417
Large – Small		-0.510***	0.020	0.229***	-0.059***	-0.777***	0.005***	0.056***
t-stat.		(-7.54)	(1.37)	(10.20)	(-9.07)	(-13.52)	(4.51)	(5.79)

Table 2: Hedge Portfolio Analysis

On June 30 of each year *t* from 2013 to 2019, we sort stocks into quintiles based on accruals, computed from either Compustat or as-filed 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 raw returns and abnormal returns of the hedging portfolios and their *t*-statistics. Panel B reports the percentage of firms in intersections between Compustat-based accruals quintiles and as-filed accruals quintiles. Panel C reports the percent 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 raw returns and abnormal returns of alternative hedging strategies. The abnormal returns are calculated using Fama-French three-factor model (i.e., *FF3 Alpha*), Carhart momentum factor model (i.e., *FF4 Alpha*), and Fama-French five-factor model (i.e., *FF5 Alpha*). Significance at the 10%, 5%, and 1% levels (two-sided) are denoted by *, **, and ***, respectively.

Panel A: Hedge Portfolio Returns

	Compustat Accruals				As-filed Accruals						
Return Measure	Q1	Q5	Hedge	*t*-stat.	Q1	Q5	Hedge	*t*-stat.	Diff. in Hedge	*t*-stat.	χ^2
Eret	1.361	1.065	0.296	1.25	1.438	0.765	0.673**	2.49	0.377**	2.46	
FF3 Alpha	0.145	-0.001	0.146	0.58	0.237	-0.324	0.561**	2.02	0.415***		10.10
FF4 Alpha	0.165	-0.004	0.169	0.68	0.245	-0.321	0.566**	2.02	0.397***		10.08
FF5 Alpha	0.182	0.008	0.174	0.72	0.270	-0.320	0.590**	2.15	0.416***		9.88

Panel B: Overlap in Accruals Quintile Ranks

	As-filed Q1	As-filed Q2	As-filed Q3	As-filed Q4	As-filed Q5
Compustat Q1	15.47%	2.01%	0.90%	0.80%	0.81%
Compustat Q2	2.78%	12.98%	2.27%	1.27%	0.72%
Compustat Q3	0.50%	3.77%	12.12%	2.59%	1.01%
Compustat Q4	0.54%	0.80%	3.75%	12.56%	2.37%
Compustat Q5	0.71%	0.44%	0.95%	2.80%	15.09%

Panel C: Transition Probabilitities

Year	Probability of Firms Staying in Compustat Q1	Probability of Firms Staying in Compustat Q5	Probability of Firms Staying in As-filed Q1	Probability of Firms Staying in As-filed Q5
2013	49.80%	38.63%	46.68%	38.59%
2014	46.32%	39.80%	47.11%	39.69%
2015	46.32%	39.39%	45.74%	39.16%
2016	46.46%	37.38%	47.64%	35.29%
2017	44.24%	38.54%	45.88%	36.15%
2018	45.93%	37.81%	44.18%	35.18%
2019	46.15%	40.43%	44.38%	37.30%
All	46.46%	38.85%	45.95%	37.32%

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:								
Long (Compustat Q1, As-filed Q1)	1.278***	2.53	0.002	0.01	0.020	0.09	0.052	0.24
Short (Compustat Q5, As-filed Q5)	0.753	1.60	-0.366**	-2.01	-0.380**	-2.07	-0.383**	-2.12
Hedge	0.525*	1.85	0.368	1.23	0.400	1.32	0.434	1.51
Compustat disagrees with as-filed:								
Long (Compustat Q1, As-filed Q2 – Q5)	0.820*	1.71	-0.269	-1.00	-0.182	-0.70	-0.237	-0.89
Short (Compustat Q5, As-filed Q1 – Q4)	1.665***	3.88	0.569***	2.87	0.544***	2.72	0.609***	3.21
Hedge	-0.844***	-2.72	-0.838**	-2.61	-0.726**	-2.37	-0.846**	-2.63
As-filed disagrees with Compustat:								
Long (As-filed Q1, Compustat Q2 – Q5)	1.412***	2.96	0.241	0.82	0.244	0.82	0.264	0.94
Short (As-filed Q5, Compustat Q2 – Q5)	0.691	1.46	-0.480**	-2.33	-0.462**	-2.21	-0.475**	-2.27
Hedge	0.721**	1.97	0.721**	1.99	0.706*	1.92	0.739**	2.07

Table 3: Fama-MacBeth Cross-sectional Regressions

This table reports the results of Fama-MacBeth cross-sectional regressions based on monthly returns from the period of July 2013 through June 2020. For each month, we estimate the following model:

$$Eret_{i,t+1} = \alpha_t + \beta_1 Accruals_{i,t} + \beta_2 Controls_{i,t} + \varepsilon_{i,t}$$

where $Eret_{i,t+1}$ denotes the excess return for firm *i* and month *t + 1*. $Accruals_{i,t}$ represents either Compustat- or as-filed accruals for firm *i* in month *t*. *Controls* include market beta (*Beta*), firm size (*Size*), market-to-book ratio (*Log(BM)*), return of month *t – 1* (*MOM_1m*), the cumulative return over the 11 months ending one month before month *t* (*MOM_12m*), the cumulative return from month *t – 36* to month *t – 13* (*MOM_36m*), asset growth (*AGR*), and cash-based operating profitability (*CbOP*). Reported coefficients and adjusted R^2 are the average values of monthly cross-sectional regressions. Fama-MacBeth *t*-statistics are reported in parentheses. All *t*-statistics are adjusted for heteroskedasticity and autocorrelations. Significance at the 10%, 5%, and 1% levels (two-sided) are denoted by *, **, and ***, respectively.

	(1)	(2)	(3)
$Accruals^{Filed}$	-2.486**		-3.611***
	(-2.33)		(-2.99)
$Accruals^{Compustat}$		0.306	3.085**
		(0.25)	(2.46)
Beta	-0.034	0.035	-0.021
	(-0.08)	(0.08)	(-0.05)
Size	0.053	0.085	0.049
	(1.13)	(1.63)	(1.06)
Log(BM)	-0.284***	-0.288***	-0.268***
	(-2.77)	(-2.80)	(-2.65)
MOM_1m	-1.745	-2.013	-1.781
	(-1.41)	(1.58)	(-1.45)
MOM_12m	0.651*	0.674*	0.660*
	(1.77)	(1.84)	(1.80)
MOM_36m	0.304*	0.323*	0.293*
	(1.69)	(1.79)	(1.66)
AGR	-0.007	0.039	0.059
	(-0.06)	(0.29)	(0.45)
CbOP	0.838*	0.886*	1.079**
	(1.81)	(1.85)	(2.19)
Intercept	-0.530	-0.840	-0.448
	(-0.68)	(-1.01)	(-0.58)
N	226,858	226,858	226,858
Adjusted R^2	0.145	0.141	0.150

Table 4: Determinants to Institutional Investors' Viewing XBRL Filings

This table reports results on the determinants of institutional investors' reliance on XBRL filings. Observations are institution-quarters. The dependent variable is an indicator variable which equals one if an institutional investor downloads any XBRL filing in the current quarter. *PriorViewing* is an indicator that equals to one if an institutional investor downloads any XBRL filing in the prior quarter. *Log (Age)* is the natural logarithm of the number of years since the institution's first appearance on Thomson Reuters. *Log(PortSize)* is the natural logarithm of the total equity portfolio size of an institution calculated as the market value of its quarter-end holdings. *Turnover* is the inter-quarter portfolio turnover rate calculated as the lesser of purchases and sales divided by the average portfolio size of the last and current quarters. *PortHHI* is the Herfindahl index of the portfolio, calculated from the market value of each component stock. *PortRet* is the monthly average return on the portfolio during the quarter. |*Flow*| is the absolute change in total portfolio value between two consecutive quarters net of the increase due to returns, expressed as a percentage of the portfolio size at the previous quarter-end. *PortVol* is the monthly portfolio return volatility during the past 12 months ending in the current quarter-end. Quarter fixed effects are included in the regression. Standard errors are adjusted for heteroskedasticity and clustering at the institution level. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels.

Dependent Variable =	*Viewing*	
	Coef.	*t*-stat.
PriorViewing	0.673***	47.09
Log(Age)	0.009***	6.54
Log(PortSize)	0.012***	11.09
Turnover	0.034***	3.58
PortHHI	0.046***	5.20
PortRet	0.001**	2.39
|*Flow*|	0.010***	2.80
PortVol	-0.059	-0.92
Year × Quarter FE	Yes	
Std. Err. Cluster	Investment Company	
N	68,759	
Pseudo R^2	0.477	

Table 5: Hedge Fund Trades and As-filed Accruals Signals

This table examines how XBRL filings downloads by a hedge fund is associated with subsequent trades made by the hedge fund. The dependent variable is *TradeValue*, which is calculated as the difference between the market value of stock holding at the end of current quarter and the market value of stock holding at the end of last quarter. If hedge fund does not hold a stock holding at the beginning or the end of current quarter, then the market value of the stock holding is assumed to be zero. *Viewing* is an indicator which equals to 1 if the hedge fund downloads the XBRL filing from EDGAR within a calendar quarter. *Accruals*^{Filed} represents as-filed operating accruals. *Diff_Accruals* is the difference between as-filed and Compustat accruals. We use the same set of controls that we used in Table 3, although for brevity the coefficients on this variables are not reported. Each column reports estimated coefficients and their *t*-statistics (in parentheses). All standard errors adjust for heteroskedasticity and are clustered at the hedge fund company level. Fund company, stock, and quarter fixed effects are included in all specifications. Significance at the 10%, 5%, and 1% levels (two-sided) are denoted by *, **, and ***, respectively.

	(1)	(2)
Viewing	0.189**	0.195***
	(2.74)	(2.85)
Accruals^{Filed}	-0.002***	
	(-6.35)	
Accruals^{Compustat}		-0.002***
		(-3.91)
Diff_Accruals		-0.003***
		(-8.97)
Viewing × *Accruals*^{Filed}	-0.217	
	(-1.11)	
Viewing × *Accruals*^{Compustat}		-0.098
		(-0.52)
Viewing × *Diff_Accruals*		-0.547**
		(-2.16)
Other Controls	Included	Included
Hedge Fund Company FE	Yes	Yes
Stock FE	Yes	Yes
Quarter FE	Yes	Yes
Standard Errors Cluster	Hedge Fund Company	Hedge Fund Company
N	119,908,720	119,908,720
Adjusted R^2	0.015	0.015

Table 6: Unrestated Compustat

On June 30 of each year *t* from 2013 to 2019, we require stocks to have non-missing values for both Compustat and as-filed accruals. However, different from the baseline analysis, Compustat operating accruals in this table are computed using *unrestated* Compustat. We sort stocks into quintiles based on operating accruals 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 quintile returns are calculated from July of year *t* to June of year *t* + 1, and the quintiles are rebalanced in June of *t* + 1. Panel A reports the average monthly raw returns and abnormal returns of the hedging portfolio and its *t*-statistics. The abnormal returns are calculated using Fama-French three-factor model (i.e., *FF3 Alpha*), Carhart momentum factor model (i.e., *FF4 Alpha),* and Fama-French five-factor model (i.e., *FF5 Alpha*). Panel B reports the results of Fama-MacBeth cross-sectional regressions based on monthly returns from the period of July 2013 through June 2020. For each month, we estimate the following model:

$$Eret_{i,t+1} = \alpha_t + \beta_1 Accruals_{i,t} + \beta_2 Controls_{i,t} + \varepsilon_{i,t}$$

where $Eret_{i,t+1}$ denotes the excess return for firm *i* and month *t + 1*. $Accruals_{i,t}$ represents either unrestated Compustat or as-filed accruals for firm *i* in month *t*. *Controls* include market beta (*Beta*), firm size (*Size*), market-to-book ratio (*Log(BM)*), return of month *t – 1* (*MOM_1m*), the cumulative return over the 11 months ending one month before month *t* (*MOM_12m*), the cumulative return from month *t – 36* to month *t – 13* (*MOM_36m*), asset growth (*AGR*), and cash-based operating profitability (*CbOP*). Reported coefficients and adjusted R^2 are the average values of monthly cross-sectional regressions. Fama-MacBeth *t*-statistics are reported in parentheses. All *t*-statistics are adjusted for heteroskedasticity and autocorrelations. Significance at the 10%, 5%, and 1% levels (two-sided) are denoted by *, **, and ***, respectively.

Panel A: Hedge Portfolio Returns

	Unrestated Compustat-Based Accruals				As-filed Accruals						
Return Measure	Q1	Q5	Hedge	*t*-stat.	Q1	Q5	Hedge	*t*-stat.	Diff. in Hedge	*t*-stat.	χ^2
Eret	1.324	0.986	0.338	1.35	1.358	0.738	0.621**	2.42	0.283*	1.74	
FF3 Alpha	0.066	-0.089	0.155	0.59	0.128	-0.338	0.516**	1.97	0.361***		7.09
FF4 Alpha	0.092	-0.102	0.194	0.74	0.133	-0.392	0.525**	1.97	0.331**		6.06
FF5 Alpha	0.119	-0.090	0.210	0.83	0.166	-0.392	0.559**	2.18	0.349**		6.39

Panel B: Fama-MacBeth Cross-sectional Regression

	(1)	(2)
Accruals^{Filed}		-2.880**
		(-2.41)
Accruals^{UR}	-1.367	0.695
	(-0.99)	(0.49)
Beta	-0.008	-0.049
	(-0.02)	(-0.12)
Size	0.077	0.052
	(1.49)	(1.11)
Log(BM)	-0.276***	-0.270***
	(-2.69)	(-2.65)
MOM_1m	-1.999	-1.798
	(-1.57)	(-1.47)
MOM_12m	0.679*	0.690*
	(1.84)	(1.88)
MOM_36m	0.300	0.298*
	(1.64)	(1.67)
AGR	0.009	-0.007
	(0.07)	(-0.06)
CbOP	0.964**	1.039**
	(2.26)	(2.19)
Intercept	-0.777	-0.514
	(-0.94)	(-0.67)
N	226,858	226,858
Adjusted *R^2*	0.142	0.150

Table 7: Potential Data Quality Issues

In Panel A, we compute an alternative measure of as-filed operating accruals after incorporating operating accruals-related custom tags. We identify operating accruals-related custom tags either by re-producing the subsection "Changes in Operating Assets and Liabilities" on the Statement of Cash Flows or through the calculation links reported by filers in their XBRL filings. In Panel B, we exclude firm-years for which the XBRL-based 10-K filings contain any operating accruals-related erroneous tags. Operating accruals-related erroneous tags are tags that are labeled as an error according to XBRL US's data quality rules. Next, as of June 30 of each year *t* over 2013 to 2019, we sort stocks into quintiles based on operating accruals 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 quintile returns are calculated from July of year *t* to June of year *t* + 1, and the quintiles are rebalanced in June of *t* + 1. Both Panel A and B report the average monthly raw returns and abnormal returns of the high-minus-low quintile (i.e., Hedge) and its *t*-statistics. The abnormal returns are calculated using Fama-French three-factor model (i.e., *FF3 Alpha*), Carhart momentum factor model (i.e., *FF4 Alpha*), and Fama-French five-factor model (i.e., *FF5 Alpha*). Panel C reports the results of Fama-MacBeth cross-sectional regressions based on monthly returns from the period of July 2013 through June 2020. For each month, we estimate the following model:

$$Eret_{i,t+1} = \alpha_t + \beta_1 Accruals_{i,t} + \beta_2 Controls_{i,t} + \varepsilon_{i,t}$$

where $Eret_{i,t+1}$ denotes the excess return for firm *i* and month *t + 1*. $Accruals_{i,t}$ represents regular Compustat ($Accruals^{Compustat}$), unrestated Compustat ($Accruals^{UR}$) or the alternative as-filed accruals ($Accruals^{Filed\_C}$) for firm *i* in month *t*. *Controls* include market beta (*Beta*), firm size (*Size*), market-to-book ratio (*Log(BM)*), return of month *t – 1* (*MOM_1m*), the cumulative return over the 11 months ending one month before month *t* (*MOM_12m*), the cumulative return from month *t – 36* to month *t – 13* (*MOM_36m*), asset growth (*AGR*), and cash-based operating profitability (*CbOP*). In Column (3) and (4), we exclude firm-years for which the XBRL-based 10-K filings contain any operating accruals-related erroneous tags. Reported coefficients and adjusted R^2 are the average values of monthly cross-sectional regressions. Fama-MacBeth *t*-statistics are reported in parentheses. All *t*-statistics are adjusted for heteroskedasticity and autocorrelations. Significance at the 10%, 5%, and 1% levels (two-sided) are denoted by *, **, and ***, respectively.

Panel A: Incorporating Accruals-Related Custom Tags

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

Panel B: Excluding XBRL Filings with Accruals-related Erroneous Tags

	Compustat-Based Accruals				As-filed Accruals						
Return Measure	Q1	Q5	Hedge	*t*-stat.	Q1	Q5	Hedge	*t*-stat.	Diff. in Hedge	*t*-stat.	χ^2
Eret	1.452	1.076	0.376	1.57	1.559	0.807	0.752**	2.63	0.376**	2.20	
FF3 Alpha	0.203	-0.014	0.217	0.85	0.283	-0.294	0.578**	1.97	0.361**		6.52
FF4 Alpha	0.226	-0.017	0.243	0.96	0.297	-0.296	0.592**	2.02	0.349**		6.25
FF5 Alpha	0.240	-0.003	0.243	0.98	0.310	-0.288	0.599**	2.08	0.356**		6.16

Panel C: Fama-MacBeth Cross-sectional Regression

	Incorporating Custom Tags	Excluding Filings with Erroneous Tags	Both
	(1)	(2)	(3)
$Accruals^{Filed\_C}$	-3.404***		-3.403***
	(-3.17)		(-3.04)
$Accruals^{Filed}$		-3.565***	
		(-2.87)	
$Accruals^{Compustat}$	2.782**	2.038**	2.768**
	(2.25)	(2.46)	(2.27)
Beta	-0.026	-0.022	-0.027
	(-0.06)	(-0.05)	(-0.06)
Size	0.048	0.067	0.066
	(1.04)	(1.38)	(1.35)
Log(BM)	-0.268***	-0.255**	-0.255**
	(-2.65)	(-2.38)	(-2.38)
MOM_1m	-1.732	-1.904	-1.844
	(-1.41)	(1.51)	(-1.46)
MOM_12m	0.658*	0.547	0.544
	(1.79)	(1.59)	(1.57)
MOM_36m	0.291	0.370*	0.368*
	(1.64)	(1.94)	(1.93)
AGR	0.044	0.061	0.046
	(0.33)	(0.45)	(0.34)
CbOP	1.080**	0.993**	0.995**
	(2.21)	(2.00)	(2.03)
Intercept	-0.445	-0.571	-0.562
	(-0.57)	(-0.73)	(-0.71)
N	226,858	201,995	201,995
Adjusted R^2	0.150	0.160	0.156

Table 8: Other Accounting-based Anomalies

In this table, we report results whether inferences drawn about other accounting-based anomalies are affected by the choice of data source. In Panel A, we report three XBRL tag attributes for each anomaly variable. *Mean Depth* is the average level of depth among XBRL tags that we used to construct each anomaly variable. The higher the level of the depth, the more disaggregated the XBRL tag is. *Max Depth* is the highest level of depth and *# Tags* is the number of tags we used to construct each anomaly variable. Additionally, we examine 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 in 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.

Panel A: Summary of Anomaly Findings

Predictor	*Mean Depth*	*Max Depth*	*# Tags*	Compustat Anomaly	As-Filed Anomaly	Difference in Anomaly
Accruals	4.00	6.00	3.00	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: XBRL Tag Attributes

Difference in Anomaly	*Mean Depth*	*Max Depth*	*# Tags*
Yes	5.41	9.20	3.20
No	3.99	5.13	2.67
Yes – No	1.42*	4.07**	0.53
(*t-stat.)*	(1.71)	(2.74)	(0.77)

Internet Appendix to

“Lost in Standardization: Revisiting Accounting-based Return Anomalies Using As-filed Financial Statement Data”

List of Supplementary Tables

Table S.1: Analysis with FactSet Financial Statement Data

In this table, we repeat our analyses in Panel B of Table 1, Panel A of Table 2 and Table 3 with FactSet financial statement data in place of Compustat data. Panel A reports the descriptive statistics on the discrepancy between the FactSet and as-filed value for each component in Hribar and Collins' (2002) formula for operating accruals. All variables are scaled by the total assets at the beginning of fiscal year. The standard errors of the two sample *t*-tests adjust for clustering at the firm and year levels. Panel B reports the average monthly raw returns and abnormal returns of the hedging portfolio and its *t*-statistics. On June 30 of each year *t* over 2013 to 2019, we require stocks to have non-missing values for both FactSet and as-filed accruals. We sort stocks into quintiles based on operating accruals 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 quintile returns are calculated from July of year *t* to June of year *t* + 1, and the quintiles are rebalanced in June of *t* + 1. The abnormal returns are calculated using Fama-French three-factor model (i.e., *FF3 Alpha*), Carhart momentum factor model (i.e., *FF4 Alpha*), and Fama-French five-factor model (i.e., *FF5 Alpha*). Panel C reports the results of Fama-MacBeth cross-sectional regressions based on monthly returns from the period of July 2013 through June 2020. For each month, we estimate the following model:

$$Eret_{i,t+1} = \alpha_t + \beta_1 Accruals_{i,t} + \beta_2 Controls_{i,t} + \varepsilon_{i,t}$$

$Eret_{i,t+1}$ denotes the excess return for firm *i* and month *t + 1*. $Accruals_{i,t}$ represents either FactSet- or as-filed operating accruals for firm *i* in month *t*. *Controls* include market beta (*Beta*), firm size (*Size*), market-to-book ratio (*Log(BM)*), return of month *t – 1* (*MOM_1m*), the cumulative return over the 11 months ending one month before month *t* (*MOM_12m*), the cumulative return from month *t – 36* to month *t – 13* (*MOM_36m*), asset growth (*AGR*), and cash-based operating profitability (*CbOP*). Reported coefficients and adjusted R^2 are the average values of monthly cross-sectional regressions. Fama-MacBeth *t*-statistics are reported in parentheses. All *t*-statistics are adjusted for heteroskedasticity and autocorrelations. Significance at the 10%, 5%, and 1% levels (two-sided) are denoted by *, **, and ***, respectively.

Panel A: Financial Statement Items Used in Calculating Accruals

	Accruals	*recch*	*invch*	*apalch*	*txach*	*aoloch*	*dpc*
Corresponding FactSet Data Item:		*ff_receiv_cf*	*ff_inven_cf*	*ff_pay_acct_cf*	*ff_pay_tax_cf*	*ff_wkcap_assets_oth*	*ff_dep_exp_cf*
FactSet Data							
Mean	-0.034	-0.010	-0.006	0.005	0.000	-0.000	0.045
Std. Dev.	0.070	0.039	0.028	0.029	0.005	0.033	0.032
Q1	-0.065	-0.019	-0.010	-0.005	0.000	-0.009	0.024
Median	-0.034	-0.004	0.000	0.002	0.000	-0.001	0.038
Q3	-0.004	0.003	0.000	0.012	0.000	0.007	0.057
As-filed Data							
Mean	-0.035	-0.009	-0.006	0.007	0.000	0.000	0.041
Std. Dev.	0.085	0.037	0.025	0.036	0.004	0.039	0.032
Q1	-0.065	-0.017	-0.009	-0.006	0.000	-0.008	0.019
Median	-0.032	-0.003	0.000	0.003	0.000	-0.000	0.035
Q3	-0.003	0.003	0.000	0.016	0.000	0.007	0.054
Two-sample Tests (As-filed Data – FactSet Data)							
Differences in mean	-0.001	0.001**	0.001***	0.002***	-0.000	0.001**	-0.004***
Clustered *t*-stat.	(-1.49)	(2.47)	(3.40)	(6.96)	(-0.36)	(2.82)	(-11.95)

Panel B: Hedge Portfolio Returns

	FactSet Accruals				As-filed Accruals						
Return Measure	Q1	Q5	Hedge	*t*-stat.	Q1	Q5	Hedge	*t*-stat.	Diff. in Hedge	*t*-stat.	χ^2
Eret	1.409	0.894	0.515**	2.12	1.491	0.748	0.743***	2.88	0.229	1.50	
FF3 Alpha	0.176	-0.266	0.442*	1.70	0.262	-0.329	0.592**	2.23	0.150		0.87
FF4 Alpha	0.190	-0.261	0.451*	1.73	0.270	-0.329	0.599**	2.24	0.148		0.82
FF5 Alpha	0.200	-0.254	0.454*	1.74	0.288	-0.331	0.619**	2.36	0.165		1.12

Panel C: Fama-MacBeth Cross-sectional Regression

	(1)	(2)	(3)
Accruals^{Filed}	-2.620**		-2.564**
	(-2.47)		(-2.46)
Accruals^{FactSet}		-1.384	0.494
		(-1.06)	(0.39)
Beta	0.022	0.049	0.028
	(0.05)	(0.12)	(0.07)
Size	0.047	0.069	0.048
	(0.98)	(1.39)	(1.01)
Log(BM)	-0.272***	-0.271***	-0.269***
	(-2.71)	(-2.72)	(-2.72)
MOM_1m	-1.782	-2.050	-1.760
	(-1.43)	(-1.59)	(-1.41)
MOM_12m	0.675*	0.687*	0.698*
	(1.80)	(1.82)	(1.87)
MOM_36m	0.286	0.293	0.275
	(1.59)	(1.56)	(1.50)
AGR	-0.033	0.019	-0.025
	(-0.21)	(0.11)	(-0.16)
CbOP	0.857*	1.038**	0.879*
	(1.75)	(2.16)	(1.75)
Intercept	-0.493	-0.733	-0.489
	(-0.62)	(-0.91)	(-0.63)
N	226,858	226,858	226,858
Adjusted R^2	0.147	0.142	0.151

Table S.2: Matched XBRL Tags for Accounting-based Return Predictors

This table presents the variable definitions in terms of the Compustat items involved for each predictor examined in Section 5.4. *L(x)* denotes the lag variable *x*.

Variable	Definition	Compustat Items	XBRL Tags (excluding the affiliated child tags)	Statement
Accruals	*(reech + invch + apalch + txach +aoloch +dpc)/L(at)*	*recch + invch + apalch + txach + aoloch*	IncreaseDecreaseInOperatingCapital	CF/S
		dpc	DepreciationDepletionAndAmortization	CF/S
		at	Assets	B/S
AGR	*at/L(at) – 1*	*at*	Assets	B/S
BM	*(seq + pstkrv)/(prcc_f × csho)*	*seq*	StockholdersEquity	B/S
		pstkrv	PreferredStockValue + PreferredStockSharesSubscribedButUnissuedSubscriptionsReceivable	B/S
CFP	*(ni + dpc)/(prcc_f × csho)*	*ni*	ProfitLoss	CF/S
		dpc	DepreciationDepletionAndAmortization	CF/S
CbOP	*2×(revt – cogs – xsga + xrd + recch + invch + apalch)/(at + L(at))*	*revt – cogs – xsga*	OperatingIncomeLoss	I/S
		xrd	ResearchAndDevelopmentExpense	I/S
		recch	IncreaseDecreaseInReceivables	CF/S
		invch	IncreaseDecreaseInInventories	CF/S
		apalch	IncreaseDecreaseInAccountsPayableAndAccruedLiabilities	CF/S
		at	Assets	B/S
dPia	*(ppegt + invt – L(ppegt) –L(invt))/ L(at)*	*ppegt*	PropertyPlantAndEquipmentGross	B/S
		invt	InventoryNet	B/S
		at	Assets	B/S
Current	*act/lct*	*act*	AssetsCurrent	B/S
		lct	LiabilitiesCurrent	B/S
DEPR	*dp/ppent*	*dp*	DepreciationAndAmortization	I/S
		ppent	PropertyPlantAndEquipmentNet	B/S
CashDebt	*2×(ib + dp)/(lt + L(lt))*	*ib*	NetIncomeLoss	I/S
		dp	DepreciationAndAmortization	I/S
		lt	Liabilities	B/S
EP	*ib/(prcc_f × csho)*	*ib*	NetIncomeLoss	CF/S
GMA	*(revt – cogs)/ L(at)*	*revt – cogs*	GrossProfit	I/S
		at	Assets	B/S

GrLtNOA	2× (Δppent + Δintan + Δao – Δlo + dp) /(at + L(at))	ppent + intan +ao	AssetsNoncurrent – LongTermInvestmentsAndReceivablesNet	B/S
		lo	LiabilitiesOtherThanLongtermDebtNoncurrent	B/S
		dp	DepreciationAndAmortization	I/S
		at	Assets	B/S
GrInvt	invt/L(invt) – 1	invt	InventoryNet	B/S
GrInvest	capx/capx_2 – 1	capx	PaymentsForProceedsFromProductiveAssets	CF/S
Lev	lt/(prcc_f × csho)	lt	Liabilities	B/S
NOA	2×((at – che) – (at – dlc – dltt – mib – pstk – ceq))/(at + L(at))/2	at	Assets	B/S
		che	CashCashEquivalentsAndShortTermInvestments	B/S
		dlc	DebtCurrent	B/S
		dltt	LongTermDebtAndCapitalLeaseObligations	B/S
		mib	RedeemableNoncontrollingInterestEquityCarryingAmount	B/S
		cep + pstk	StockholdersEquity	B/S
OP	2× (revt – cogs – xsga + xrd) /(at + L(at))/2	revt – cogs – xsga	OperatingIncomeLoss	I/S
		xrd	ResearchAndDevelopmentExpense	I/S
		at	Assets	B/S
Quick	(act – invt)/lct	act	AssetsCurrent	B/S
		lct	LiabilitiesCurrent	B/S
		invt	InventoryNet	B/S
RealEstate	(fatb + fatl)/ppegt	fatb	BuildingsAndImprovementsGross + ConstructionInProgressGross + LeaseholdImprovementsGross + PropertySubjectToOrAvailableForOperatingLeaseGross	B/S
		fatl	CapitalLeasedAssetsGross	B/S
		ppegt	PropertyPlantAndEquipmentGross	B/S
TB	(txfo + txfed)/(0.35×ib)	txfo + txfed	CurrentIncomeTaxExpenseBenefit	I/S
		ib	NetIncomeLoss	I/S

Table S.3: Other Anomalies with Significant Discrepancy between Two Data Sources

Panel A: Earning Before Depreciation and Extraordinary Items-to-Total Debt (*CashDebt*)

	Compustat				As-filed						
Return Measure	Q1	Q5	Hedge	*t*-stat.	Q1	Q5	Hedge	*t*-stat.	Diff. in Hedge	*t*-stat.	χ^2
Eret	0.744	1.039	0.295	0.53	0.549	1.174	0.624	0.99	0.330	1.43	
FF3 Alpha	-0.488	0.023	0.511	1.27	-0.838	0.145	0.983**	2.25	0.472**		4.12
FF4 Alpha	-0.488	0.029	0.517	1.27	-0.833	0.145	0.978**	2.22	0.461**		4.05
FF5 Alpha	-0.354	0.015	0.369	1.13	-0.683	0.138	0.821**	2.38	0.452**		4.65

Panel B: Growth in long-term net operating assets (*GrLtNOA*)

	Compustat				As-filed						
Return Measure	Q1	Q5	Hedge	*t*-stat.	Q1	Q5	Hedge	*t*-stat.	Diff. in Hedge	*t*-stat.	χ^2
Eret	0.785	0.919	0.164	0.59	0.475	1.288	0.812***	3.16	0.648**	2.64	
FF3 Alpha	0.058	-0.064	-0.122	-0.43	-0.305	0.237	0.541**	2.04	0.663***		6.77
FF4 Alpha	0.062	-0.059	-0.121	-0.43	-0.290	0.236	0.527**	2.02	0.648**		6.49
FF5 Alpha	0.020	-0.025	-0.045	-0.18	-0.343	0.274	0.617**	2.57	0.662***		7.22

Panel C: Operating Profitability (*OP*)

	Compustat				As-filed						
Return Measure	Q1	Q5	Hedge	*t*-stat.	Q1	Q5	Hedge	*t*-stat.	Diff. in Hedge	*t*-stat.	χ^2
Eret	0.003	1.359	1.355**	2.36	0.269	1.132	0.863**	2.07	-0.492*	-1.77	
FF3 Alpha	-1.153	0.251	1.404***	3.36	-0.646	0.154	0.800**	2.23	-0.604*		2.65
FF4 Alpha	-1.144	0.251	1.395***	3.33	-0.623	0.155	0.778**	2.22	-0.607*		2.75
FF5 Alpha	-1.051	0.252	1.303***	3.54	-0.674	0.149	0.823**	2.74	-0.480*		2.61

Panel D: Taxable Income to Book Income (*TB*)

	Compustat				As-filed						
Return Measure	Q1	Q5	Hedge	*t*-stat.	Q1	Q5	Hedge	*t*-stat.	Diff. in Hedge	*t*-stat.	χ^2
Eret	0.988	1.557	0.569**	2.09	0.919	1.713	0.794***	2.92	0.188**	2.02	
FF3 Alpha	-0.380	0.306	0.686**	2.18	-0.452	0.466	0.918***	3.37	0.232*		2.91
FF4 Alpha	-0.375	0.301	0.677**	2.24	-0.439	0.459	0.898***	3.33	0.221*		2.68
FF5 Alpha	-0.372	0.319	0.690**	2.19	-0.455	0.475	0.920***	3.36	0.230*		2.94