Textual Analysis and International Financial Reporting:

Large Sample Evidence

Mark Lang†*

Lorien Stice-Lawrence†

October 2014

We examine annual report text for over 15,000 non-US companies from 42 countries over the period 1998-2011, focusing on the length of disclosure, presence of boilerplate, comparability with US and non-US firms, and complexity. We find that textual attributes are predictably associated with regulation and incentives for more transparent disclosure and are correlated with economic outcomes such as liquidity, institutional ownership, and analyst following. Using mandatory IFRS adoption as an exogenous shock, annual report disclosure improved in the sense that quantity of disclosure increased, boilerplate was reduced, and comparability increased relative to both US and non-US firms. Firms with the greatest improvements in financial reporting experienced the greatest improvements in economic outcomes around IFRS adoption.

Textual Analysis and International Financial Reporting:

Large Sample Evidence

October 2014

We examine annual report text for over 15,000 non-US companies from 42 countries over the period 1998-2011, focusing on the length of disclosure, presence of boilerplate, comparability with US and non-US firms, and complexity. We find that textual attributes are predictably associated with regulation and incentives for more transparent disclosure and are correlated with economic outcomes such as liquidity, institutional ownership, and analyst following. Using mandatory IFRS adoption as an exogenous shock, annual report disclosure improved in the sense that quantity of disclosure increased, boilerplate was reduced, and comparability increased relative to both US and non-US firms. Firms with the greatest improvements in financial reporting experienced the greatest improvements in economic outcomes around IFRS adoption.

I.   Introduction

A substantial and growing literature examines and compares the characteristics of accounting data, particularly earnings, for firms worldwide. However, much less is known about the associated textual disclosure, owing in part to the difficulty in obtaining full text annual reports and in part to the difficulty in quantifying textual data.

Our goal in this paper is to provide what is, to our knowledge, the first large-sample empirical analysis of annual report textual disclosure for over 15,000 non-U.S. firms from 42 countries over the period 1998-2011. We apply textual analysis to characterize disclosure across a number of dimensions: quantity of disclosure as measured by annual report length, use of boilerplate disclosure as measured by the use of standardized discussion, comparability as measured by the similarity of disclosure across firms within an industry relative to both US and non-US firms, and complexity of disclosure as measured by the Gunning Fog index.

We conduct analyses to understand cross-sectional determinants of disclosure attributes as well as trends over time. We also focus on one specific event, IFRS adoption, using a difference-in-differences design to assess the effects of an exogenous shock on disclosure. We correlate disclosure attributes with economic "outcomes" both in time series and cross-section, as well as around IFRS adoption, to assess how aspects of disclosure are associated with factors such as liquidity, analyst following, and institutional ownership. We focus on a broad set of firms, countries, and years to provide as complete a picture as possible of the characteristics, trends, and economic outcomes associated with textual disclosure.

From a descriptive perspective, we provide evidence of substantial cross-country variation in textual reporting. Perhaps most striking, though, are trends over time. The quantity of textual reporting has increased substantially. In addition, comparability has improved, both

1

among non-US reports as well as between non-US and US reports. Boilerplate increased during the first half of the sample but has decreased in more recent years. Fog has generally increased over the sample period, perhaps reflecting the increase in disclosure of more complex topics (e.g., financial instruments, currencies, taxes, and share-based compensation).

In terms of determinants, and controlling for a variety of other factors, our textual attributes appear to be significantly influenced by regulation and the demand for information. Annual report disclosure is significantly longer for firms applying IFRS or US GAAP and firms which are audited by Big-5 auditors. In addition, disclosure for IFRS and US GAAP firms contains less boilerplate and tends to be more comparable with both US and non-US firms. The analysis for IFRS and US GAAP is robust to inclusion of firm and year fixed effects, suggesting that differences in firm-level economics or time periods do not drive the results.[1]

Next, we correlate our disclosure measures with liquidity, institutional ownership, and analyst following. There is intuitive appeal and theoretical support for the notion that longer annual reports with less boilerplate and greater comparability should be correlated with greater transparency leading to greater liquidity, analyst following, and institutional ownership. However, given that our constructs are, by their very nature, imprecise and there is little existing literature applying them, we want to establish that they do, indeed, correlate with informational variables. For parsimony, we first aggregate our textual variables using factor analysis and then assess whether those aggregated measures are correlated with economic outcomes.

---

[1] We take several steps to ensure that our comparisons are "apples to apples." First, the fact that results are robust to firm and year fixed effects suggests that we are not capturing country-level, firm-level or time period idiosyncrasies. Second, the reports are checked by the data provider to ensure basic consistency. Third, the data analysis software we apply eliminates tabular and other information which appears inconsistent with English textual disclosure. Fourth, we also exclude reports that are US regulatory filings. Finally, we hand check a sample of cases which appear to be unusual in length or content to ensure that these procedures are effective.

We find that liquidity, institutional ownership, and analyst following tend to be higher for firms with greater lagged quantities of disclosure, less boilerplate, and greater comparability with US firms. While we are hesitant to draw strong conclusions on causality, results are robust to inclusion of firm fixed effects and lagged independent variables, as well as an instrumental variables analysis, suggesting that increases in our annual report variables tend to precede increases in liquidity, institutional ownership, and analyst following.

Next, we investigate one particular innovation in more detail: the mandatory adoption of IFRS. IFRS adoption provides an interesting context for at least two reasons. First, from the firm's perspective, it represents a significant exogenous shock to accounting standards and regulation. One of the difficulties with disclosure research in general is the fact that it suffers from potential endogeneity concerns. However, mandatory IFRS adoption was largely outside of the firm's control. Second, and more importantly, while there is a substantial body of research examining the characteristics of accounting information around IFRS adoption, we know much less about its effect on textual disclosure. Using a difference-in-differences design, we document striking changes in disclosure around IFRS adoption. Relative to our control sample, IFRS adoption increased the "quality" of annual reports in the sense that it increased the amount of disclosure, as well as decreasing repetitive disclosure and increasing similarity with both US and non-US peer firms. Using Latent Dirichlet Allocation (LDA) analysis to examine how IFRS adoption affected the *topics* that a firm chooses to disclose in its annual report, we find that these overall increases in disclosure length around adoption are driven by increases in disclosure about, for example, financial instruments, compensation, accounting policies, and general financial performance. It is notable that, while empirical evidence on the impact of IFRS with respect to quantitative data is mixed (see, for example, Barth et al. (2008) and Ahmed et al.

(2013)), the effects on textual characteristics are striking, suggesting that the impact of IFRS on textual disclosure may have been at least as pronounced as its effect on measurement.

Finally, we correlate changes in our measures of textual disclosure with economic outcomes around IFRS adoption. We first use mandatory IFRS adoption as well as peer disclosure variables as instruments for firm-level disclosure in an instrumental variables analysis to document that changes in disclosure are associated with economic outcomes. We also examine variation in the extent to which firms benefitted from mandatory IFRS adoption by showing that firms that experienced the largest increases in the quantity and comparability of disclosure also enjoyed greater increases in liquidity, institutional ownership, and analyst following. Firms that experienced increases in fog and boilerplate showed little or no benefit of IFRS adoption.

Our analysis is subject to several caveats. First, our approach and methods are somewhat exploratory because development of textual measures, especially in the international context, is still at an early stage which, in some cases, means that we develop our own metrics. We focus on measures that we believe are likely to be of importance to regulators and investors globally (e.g., quantity of disclosure, comparability, boilerplate, and complexity). While we take some comfort in the fact that our textual measures correlate with determinant and outcome variables in predictable ways, and change notably around IFRS adoption, we recognize that there is necessarily noise in our measurement. That being said, we believe that it is important to explore approaches for quantifying global textual data, appropriately caveated.

Second, we adopt a large sample approach. There are advantages and disadvantages to a broad-brush approach relative to one that focuses on specific samples or selections of words. While our approach inevitably ignores some of the subtleties of disclosure that might be captured

4

by a more focused analysis, it allows us to take a first step in examining textual disclosure choices more broadly for a wide sample of firms and relatively long time series.

Third, we examine only English-language disclosures. While English-language disclosures are important given their prevalence in modern commerce, our approach excludes local language disclosure and firms that report only in non-English languages. There is scope for additional research incorporating non-English text to address international disclosure more broadly.[2]

In the next section, we discuss the related literature. Then, we describe our measures, hypotheses, empirical design, and results. Finally, we provide conclusions and potential directions for future research.

II.   Related Literature

Our research is related to at least two streams of literature. The first is the international research on the characteristics of accounting information. Beginning with papers such as Alford et al. (1993), researchers have examined the characteristics of accounting data, particularly net income and shareholders' equity, across countries. Research such as Land and Lang (2002) investigates harmonization trends in accounting data over time. More recently, several papers have investigated the effects of IFRS adoption on the characteristics of accounting data. While some papers suggest that accounting quality has improved with IFRS adoption (e.g., Barth et al., 2008), others provide more mixed evidence (e.g., Ahmed et al., 2013). Although understanding the characteristics of accounting data is clearly important, our goal is to provide a broader assessment of the more complete reporting package. Arguably, the associated textual disclosure is at least as important as summary accounting measurement.

---

[2] See Jeanjean et al. (2010) for an analysis of which firms choose to report in English and Jeanjean et al. (2014) for the effects of choosing to report in English. We evaluate the sample selection effects of requiring English disclosure in our empirical analyses.

Second, our research is related to the growing body of papers applying textual analysis, primarily in the US. For example, Li (2008) measures the readability of annual reports using the Gunning Fog index and finds that firms with poor readability have lower current and future performance and earnings persistence, while Miller (2010) and Lehavy et al. (2011) provide evidence that readability affects investors and analysts, respectively. Other papers examine Fog and textual attributes such as tone for US firms (see Li 2010 for examples). Outside of the US, there is little broad sample empirical evidence on textual annual report disclosure.[3]

We are not aware of research that examines textual annual report content for large samples of non-US firms over significant time periods or around significant accounting changes such as IFRS adoption. Given the important cross-sectional variation in institutions and incentives, as well as changes over time, our goal is to provide evidence on textual characteristics in a global setting. In addition, we develop and validate textual disclosure measures such as boilerplate and comparability which we believe have the potential to contribute to the literature going forward.

III.   Primary Textual Constructs

There are a variety of textual attributes we could consider to characterize annual report disclosure. We focus on five that we believe are likely to be of interest to regulators and investors, and which have the potential to affect the information content of the annual report.

---

[3] Bischof and Daske (2013) examine sovereign risk disclosures for a sample of European banks, and Ernstberger and Grüning (2013) examine a single year of disclosure for a sample of European firms using artificial intelligence techniques. Also related, papers such as Brochet et al. (2013) study the impact of language distance on the complexity of conference call transcripts and Lundholm et al. (2013) examine use of numbers and textual clarity for foreign firms trading on US exchanges relative to US firms.

a. Annual Report Length

First, we consider the length of the disclosure, as measured by the number of words in textual sections of the annual report. All else equal, we expect longer annual reports to be more informative. We measure annual report length as the natural log of the total number of words contained in the annual report. Because we also develop a measure of boilerplate disclosure, the measure of length that we focus on is the length of the *non-boilerplate* portion of the text, LN_NBWORDS, but inferences are unchanged if we use the length of the entire document.

b. Boilerplate

Beyond the quantity of disclosure, one particular attribute that has been identified by regulators and standard setters as problematic is the use of "boilerplate" in annual reports. For example, Hans Hoogervost, chairman of the IASB, has identified boilerplate as a primary concern for standard setters, and notes that increasing the quantity of disclosure is not helpful if it simply reflects disclosure that is meant to, for example, reduce legal or reputational exposure rather than communicating additional information. He adds that boilerplate may even provide opportunities to hide information, reducing overall informativeness (Hoogervost, 2013). The FASB has also identified boilerplate as a frequent concern with the content of the annual report (FASB, 2012). Consistent with the view of the IASB and FASB, all else equal, we expect annual reports containing more boilerplate to be less informative.

We define boilerplate as standardized disclosure that is so prevalent that it is unlikely to be informative. We measure boilerplate by identifying 4-word phrases (tetragrams) that are extremely common among documents in a firm's home country and measuring the extent to

which they appear in a given firm's annual report.[4] Our measure relies on the assumption that

the use of extremely common phrases is a sign of boilerplate because disclosure that

indiscriminately duplicates the disclosure of many other firms is unlikely to contain important

firm-specific information. Our measure, BOILERPLATE, is the percent of words in the annual

report from sentences that contain at least one of these "boilerplate" phrases.[5] See the Appendix

for additional details on how we compute this measure, including how we exclude phrases that

are likely to be part of mandatory regulatory disclosure.

c. Comparability

Standard setters, regulators, and other stakeholders have long argued that comparability

across, for example, firms within an industry is an important attribute for users of financial

statement information. The IASB and FASB concept statements identify comparability as one of

four characteristics that enhance the usefulness of information that is relevant and faithfully

represented, and one of the primary goals of the IASB has been to improve the comparability of

accounting information globally.

True comparability is difficult to measure because it goes beyond similarity or

uniformity.[6] In particular, it is difficult to disentangle whether similarity is "making unlike

things look alike." We focus on the similarity of textual information because it is easier to

operationalize and attempt to compare "like" things by comparing disclosure within industry

---

[4] Barrett et al. (2006) discuss why the distributional properties of word phrases make it very unlikely that different authors would use identical phrases, even when discussing similar topics, and researchers have exploited this fact to identify duplication such as plagiarism.

[5] Because our boilerplate measure is calculated as a percent of total word count, it may be affected by document length, where the same number of repeated phrases makes up a smaller proportion of a longer document. To control for mechanical effects, we include powers of total word count as control variables in BOILERPLATE regressions, similar to Brown and Tucker (2011). In addition, sentences containing boilerplate could contain useful numerical information which makes them informative. Inferences are unchanged if we identify informative numbers using the approach in Blankespoor (2014) and exclude boilerplate sentences that include these numbers.

[6] In Concepts Statement No. 8, the FASB clarifies that, "Comparability is not uniformity. For information to be comparable, like things must look alike and different things must look different. Comparability of financial information is not enhanced by making unlike things look alike any more than it is enhanced by making like things look different." (FASB, 2010)

groupings, recognizing that this approach is subject to caveats. We focus on two comparison groups, non-US firms and US firms, because a primary goal of the IASB has been to increase comparability in reporting among non-US firms, and one of the goals of the convergence effort between the IASB and FASB has been to increase comparability between US and non-US firms. These two splits are also of interest because they permit us to assess whether it matters with whom a firm is similar.

Both of our measures of comparability are estimated by comparing how similar the words a firm uses in its annual report are with those used by its non-US and US peers. We compare reports using cosine similarity, a textual measure which identifies similar documents by comparing the relative word frequencies across documents. It is bounded between 0 and 1, where documents with identical proportions of words have a score of 1, and those with no overlapping words have a score of 0. Several recent studies in accounting and finance have used variants of this measure. For example, Brown and Tucker (2011) examine year-to-year changes in MD&A, Peterson et al. (2012) measure the similarity in firms' descriptions of accounting practices, and Hoberg and Phillips (2010) compare product descriptions in 10-K filings.

We compute the comparability of each firm as the median of its cosine similarity with all of its current peer firms, where non-US peers are used to calculate XUS_COMP and US peers are used to calculate US_COMP. Because we are interested in capturing comparability (i.e., similar underlying economics being reported similarly), we match our sample firms to non-US and US peers in the same 3-digit ICB (Industrial Classification Benchmark) industry and fiscal year, and choose the 30 firms with the closest matches of lagged total assets. Additional details about the comparability measures are given in the Appendix.

d. Fog

Our final measure is Fog. Papers such as Li (2008) use the Gunning Fog Index to measure the readability of text based on the length of sentences and the proportion of complex terminology. We include Fog because it has been the focus of much of the research on textual attributes in the US. However, interpretation is clouded in our context by the fact that Fog likely combines the complexity of the underlying economics with the complexity of disclosure. For example, an increase in the disclosure of financial instruments and hedging will, of necessity, likely increase the complexity of disclosure. Nevertheless, we view Fog as a potentially important construct given the prior literature. Fog is defined as:

$$Fog = (words\_per\_sentence + percent\_complex\_words) * 0.4$$

where complex words are words with three or more syllables.[7]

IV. Hypotheses:

Given the lack of prior research on textual attributes, especially in international contexts, our investigation is to some extent descriptive. However, we develop our analyses around general hypotheses linking our textual measures to determinants and outcomes.

In our first set of analyses, we investigate the link between our textual attributes and factors such as accounting standards (US GAAP and IFRS), auditor quality (large vs. small auditors), and exchange listings (ADR's on US markets). For parsimony, we refer to IFRS and US GAAP as "non-local accounting standards" and auditor quality and ADR listing as "oversight," although we acknowledge that these variables have broader effects.

---

[7] The Fog index can be interpreted as the number of years of education an average person would need to understand a given piece of writing. For example, the children's book "The Cat in the Hat" has a Fog index of 3 whereas the Declaration of Independence has a Fog index of 20. See Bushee et al. (2014) for a discussion of the difficulty in interpreting linguistic complexity in the context of financial reporting.

Hypothesis 1: Non-local accounting standards and greater oversight

> *a. increase the quantity of textual disclosure,*

> *b: increase the comparability of disclosure among non-US firms,*

> *c: increase the comparability of disclosure with US firms,*

> *d: decrease boilerplate, and*

> *f: affect the level of Fog.*

We predict that the quantity of disclosure will be higher for firms with non-local accounting standards based on the notion that US GAAP and IFRS impose (relative to local GAAP) more detailed disclosure requirements. We also predict that higher quality auditors and SEC oversight result in firms complying more fully with required disclosure.

We predict that adoption of non-local accounting standards, as well as large auditors and SEC oversight, will increase the comparability of reporting among non-US firms as they promote more standardized disclosure. However, because our comparison firms include both local GAAP and non-local GAAP firms, the extent to which nonlocal standards increase comparability overall is an empirical question.[8]

We expect US GAAP and US listing to increase the comparability between non-US firms and US firms because of shared accounting standards and regulatory requirements. IFRS will increase the comparability between non-US firms and US firms to the extent that IFRS has converged towards US standards.[9] Similarly, if large auditors enhance the consistency of reporting (and given the fact that the Big-5 originated in the US), we expect greater comparability with US firms for firms employing large auditors.

---

[8] We also examine comparability among IFRS firms only and demonstrate that, as expected, IFRS adoption has stronger comparability-increasing effects among IFRS firms than between IFRS firms and local GAAP firms.
[9] We also predict, and find evidence, that US GAAP has more of an effect on comparability with US firms, while IFRS has more of an effect on comparability among non-US firms.

Given the focus of the FASB, SEC, and IASB on reducing boilerplate, we expect adoption of IFRS or US GAAP, or listing on US exchanges, to reduce boilerplate. Similarly, to the extent that large auditors focus on the quality of disclosure, we expect lower levels of boilerplate. However, it is an open question as to whether boilerplate has actually been reduced in practice.

Our hypothesis with respect to Fog is non-directional because it can be driven by both beneficial and detrimental aspects of disclosure. Prior research provides evidence that Fog may be linked to managerial obfuscation (Li 2008) and higher processing costs (Miller 2010), but it is affected by multisyllabic words which can be associated with informative technical disclosure. If nonlocal accounting standards and oversight simplify language, we expect them to reduce Fog. However, they may necessitate increased discussion of technical topics, increasing Fog.

Hypothesis 2: Increased quantity and comparability of disclosure are associated with greater liquidity, analyst following, and institutional ownership.

For parsimony in our outcomes investigation, we begin by conducting a factor analysis of textual attributes. The first factor is positively associated with the quantity of disclosure and comparability of disclosure with both non-US and US firms. We predict that this factor will be positively associated with liquidity, analyst following, and institutional ownership for two reasons. First, more disclosure, all else equal, should increase informativeness and decrease information asymmetry among investors.[10] Second, Defranco et al. (2011) argue that comparability enhances the ability of investors and other financial statement users to understand and predict economic events by decreasing information acquisition and processing costs and by providing information about shared economic factors.

---

[10] Prior research suggests that liquidity (Glosten and Milgrom (1985)), analyst following (Lang and Lundholm (1996)) and institutional ownership (Bushee and Noe (2000)) are higher for firms with higher levels of disclosure.

Hypothesis 3: Increased fog and boilerplate are associated with reduced liquidity, analyst following, and institutional ownership.

Our second factor is positively associated with boilerplate and Fog, and we expect it to be negatively correlated with liquidity, analyst following, and institutional ownership. Conditional on the quantity of disclosure, we expect lower information content in disclosure that is characterized by high levels of boilerplate (Hoogervost (2013)) and low levels of readability (Li (2008)). We also expect both boilerplate and low readability to increase the difficulty for financial statement users to extract the available information from the financial statements, increasing their information acquisition costs.

Our last hypothesis examines textual characteristics in the context of IFRS adoption and is presented in two parts:

Hypothesis 4.a: Mandatory IFRS adoption is associated with increases in the quantity of textual disclosure, increases in comparability among non-US firms and with US firms, reductions in boilerplate, and changes in Fog.

This hypothesis follows from the discussion of IFRS in Hypothesis 1. While the specific financial reporting effects associated with IFRS adoption likely varied based on firms' particular circumstances, the effect on the typical firm was substantial.[11] Although firms faced disclosure requirements in the local environment, we expect that mandatory adoption of IFRS increased the quantity of disclosure required in most jurisdictions.

In addition, the fact that IFRS provided a shared set of guidance relative to local GAAP likely increased the comparability of reporting among non-US firms (particularly those that

---

[11] For example, IFRS requires substantive disclosure regarding financial instruments (IFRS 7), employee benefits and pensions (IAS19), share-based payments (IFRS 2), operating segments (IFRS 3), income taxes (IAS 12), provisions and contingencies (IAS 37), property, plant and equipment (IAS 16), intangible assets (IAS 38), foreign exchange rates (IAS 21), asset impairments (IAS 36), related party transactions (IAS 24), and leases (IAS 17).

13

adopted IFRS) and with US firms due to convergence between the two sets of standards. We also expect a reduction in boilerplate given the fact that IFRS was designed to provide substantive disclosure, and many firms were coming from environments in which disclosure had traditionally been influenced by statutory requirements. We do not have strong predictions for Fog because, while IFRS might have provided clearer disclosure relative to local requirements, it also likely increased disclosure of technical topics.

Hypothesis 4.b: Firms with the largest increases in desirable disclosure attributes around mandatory IFRS adoption (increases in the quantity and comparability of disclosure) experienced greater increases in liquidity, analyst following, and institutional investment, while those with the largest increases in undesirable disclosure attributes (increases in boilerplate and Fog) experienced smaller benefits of IFRS adoption.

This prediction follows from hypotheses 2 and 3 and is related to prior studies examining variation in benefits from IFRS adoption. If the associations we document in the overall sample reflect the effect of accounting standards and oversight, we expect the effects of IFRS on liquidity, analyst following, and institutional investment to be most pronounced in contexts in which mandatory adoption was associated with the greatest increases in the quantity and comparability of reporting, and to be mitigated for firms with increases in boilerplate and Fog.[12]

V.   Sample Selection

We gathered annual reports from the Global Reports database in Bureau van Dijk's Osiris, which contains annual reports for public companies in more than 100 countries beginning in the 1990s. We restrict the sample to the 42 countries that had at least 1,000 annual reports as

---

[12] The notion that the benefits of IFRS adoption vary cross-sectionally is consistent with research such as Daske et al (2013) which focuses on "serious" adopters and Christensen et al. (2013) which focuses on enforcement.

of September, 2012.[13] Osiris acquires annual reports directly from the company or through the

company's website. Of the resulting 194,973 English reports, we retained 87,608 unique reports

which had the appropriate accounting data and passed our data screens.[14]

Table 1 presents descriptive data by country and year. Observations are spread across a

wide range of countries including emerging market economies. The largest number of

observations comes from the UK which comprises approximately 15% of the sample. The

number of observations increases over time based on availability of electronic annual reports,

with the largest number in 2009, but there are a substantial number of observations throughout

the 2000's. We include firm and year fixed effects in our primary analyses so results should not

be unduly influenced by changes in sample composition over time.

Table 2, Panel A provides descriptive statistics for the primary variables in the paper.

The median firm is fairly large, with total market capitalization of $114.5 million, but the

interquartile range extends from $24.7 to $588.9 million. About a third of the observations are

reported under IFRS, and 2% apply US GAAP or trade on US exchanges. About half of the

firms are audited by Big-5 affiliated auditors. The textual component of the average annual

report has about 17,000 words, and the interquartile range is from 11,000 to 26,000 words.

Table 2, Panels B and C provide descriptive statistics comparing our sample to the

Datastream population overall and split by English and non-English speaking countries. Our

sample includes about a third of the Datastream population, but is more representative of firms

with analyst data and firms in English-speaking countries, suggesting that Osiris was more likely

---

[13] Our data end in fiscal 2011 because Osiris no longer provides academic access to full text financial reports. We did not include reports for companies listed under the Cayman Islands because most of these companies are domiciled in other countries.

[14] See the Appendix for details of how we parsed the data and screened for errors and duplicates.

to include firms followed by analysts and firms in English-speaking countries. In our analyses, we discuss results comparing English and non-English countries.

Table 2, Panel D reports descriptive statistics for our textual measures over time.[15] Despite the fact that the average size of the sample firms has decreased over time due to increases in the breadth of the sample, the number of words in the typical annual report has increased. In general, Fog has increased over time, perhaps reflecting increased disclosure of complex topics. Boilerplate increased during the pre-IFRS period, but has decreased since 2004. Perhaps most interesting, reporting has become consistently more similar over time, both relative to the US as well as relative to other non-US firms. The increase in similarity has been more pronounced among non-US firms, likely reflecting the effects of IFRS in converging accounting practice explored in later analyses. Especially following 2005, the similarity among non-US firms increased relative to the similarity with US firms, again likely reflecting the effect of IFRS.

Table 2, Panel E presents descriptive statistics on textual attributes by country, after controlling for size and year. While we do not perform formal statistical analyses, Asian countries, especially China, Hong Kong, Singapore, and Indonesia, tend to have higher Fog and Boilerplate scores. In contrast, Fog and Boilerplate tend to be lower in European countries, especially the Netherlands, Norway, and Sweden.

Results by industry (not tabulated) suggest that, controlling for other factors, disclosure tends to be particularly lengthy for banks and insurance companies but shorter for industrial companies such as automobile manufacturers and chemicals. Interestingly, bank and insurance disclosure also tends to be more comparable, both among non-US firms as well as with US firms,

---

[15] To ensure that the patterns in Table 2, Panel D do not reflect changes in sample composition over time, we also investigated trends after controlling for a wide array of fundamentals such as size, year, industry, profitability, leverage, and Big-5 auditor. Inferences are robust to inclusion of controls, suggesting that changes in sample composition do not drive the observed trends.

perhaps reflecting additional oversight in the banking sector and the idiosyncratic nature of banking relative to other industries. Fog tends to be high in insurance, telecommunications, and healthcare, likely reflecting the prevalence of technical language in those industries.

VI.   Correlations

Table 3 presents correlations. Several sets of comparisons are relevant. First, the correlations among the various textual attribute measures are not particularly high, with most correlations below about 30%. The exception is the correlation between US similarity and non-US similarity (52%), indicating that annual reports that are more similar to non-US peer firms also tend to be more similar to US peers. The other correlations are generally consistent with intuition. Fog tends to be positively correlated with percent boilerplate (21%), while percent boilerplate is negatively correlated with non-boilerplate annual report length (-11%).

Second, the correlations between the textual variables and our primary variables of interest give a sense for the results to follow, albeit without controls. IFRS is positively correlated with similarity among non-US firms (36%) and with US firms (15%), consistent with the goal of the IASB to increase comparability of reporting among non-US firms and to converge with US GAAP. IFRS, US GAAP, Big-5 Auditor, and ADR have correlations with our textual characteristics consistent with the hypothesis that they lead to higher reporting quality in terms of longer disclosure, less boilerplate, and greater comparability with non-US and US firms.

Finally, the correlations provide preliminary evidence on the relation between the textual variables and various market outcomes generally consistent with our expectations. Liquidity, analyst following, and institutional ownership tend to be greater for firms whose annual reports

17

are longer, contain less boilerplate and Fog, and which are more comparable to US firms.[16]

These correlations are reassuring because they suggest that market outcomes are correlated with our text-based measures in predictable ways. While these correlations do not control for other factors, regression results going forward which include numerous controls, as well as firm and year fixed effects, yield similar inferences.[17]

VII. Regression Results

Table 4 provides our first set of regression results examining determinants of the textual characteristics. We first present results of a mixed model with random firm effects and fixed year effects in Panel A. While we include firm and year fixed effects in our remaining analyses, applying random firm effects in Panel A permits us to investigate the impact of variables which exhibit very little within-firm variation over our sample period, such as Big-5 auditor and ADR.

In Panel B, we include firm fixed effects so that our analyses going forward are essentially within-firm comparisons. Regression results are generally consistent with the correlations in Table 3, indicating that conclusions are robust to less restrictive specifications permitting cross-firm variation.[18] We include year fixed effects to abstract from trends in textual attributes over time as well as from changes in general macroeconomic conditions. Standard errors throughout are clustered by firm.

---

[16] We also considered the measure of liquidity based on price pressure from Amihud (2002): absolute value of returns divided by volume. That measure is highly correlated with zero return days and bid/ask spreads, and yields similar significant correlations with our primary variables of interest.

[17] In addition, we lose observations due to requirements related to control variables in the regression specifications. The fact that the regression results are consistent with the correlations suggests that regression results are likely to be representative of the broader sample.

[18] Results are robust to other approaches such as industry and country fixed effects. We also estimated our regressions including controls for earnings quality based on Wysocki (2008). We exclude them from our reported results because they significantly reduced the sample size and did not alter our conclusions.

Because our initial goal is to understand the determinants of the textual attributes, particularly those related to aspects of oversight, regulatory environment, and incentives, we control for a wide range of potential economic determinants. Annual report length is positively associated with firm size, book/market ratio, and earnings surprise. The notion that larger firms would have more to disclose is not surprising. If book/market reflects firm maturity and stability, greater disclosure is likely driven by greater underlying complexity. Large earnings surprises also likely necessitate greater disclosure. Boilerplate tends to be lower for firms that are larger and those reporting losses, suggesting that greater uncertainty necessitates less generic disclosure. Fog is higher for smaller firms and for firms that are less profitable or reporting losses, consistent with Li (2008).

In terms of the primary relations of interest, IFRS is strongly associated with each of our individual textual measures in both specifications. IFRS annual reports tend to be significantly longer than non-IFRS annual reports, consistent with Hypothesis 1.a and with IFRS increasing overall disclosure requirements. IFRS reports exhibit greater similarity with other non-US annual reports, consistent with Hypothesis 1.b. Existing research provides mixed evidence on the extent to which IFRS reporting increases comparability of accounting amounts such as net income among firms (e.g., Yip and Young 2012 and Cascino and Gassen 2012). However, our results suggest that the effects of IFRS on textual comparability are quite strong. In addition, IFRS reports exhibit greater similarity with US annual reports, consistent with Hypothesis 1.c and with the evidence in Barth et al. (2012) that IFRS increased measurement comparability with US firms, and suggesting that the efforts toward convergence were manifested not only in measurement but in textual disclosure as well. Boilerplate is significantly lower for IFRS firms, suggesting that the additional length did not also increase boilerplate, consistent with Hypothesis

19

1.d. Overall, the effect of IFRS on textual reporting on the whole appears to be "good" in the sense that it is associated with more disclosure, less boilerplate, and greater similarity with both US and non-US firms.

US GAAP generally has similar effects as IFRS in the sense that it is associated with more textual disclosure, less boilerplate, and greater similarity with both US firms and non-US firm, consistent with Hypotheses 1 a-d. As expected, the effect of US GAAP is stronger for comparability relative to US firms than comparability relative to non-US firms while IFRS is more strongly associated with comparability relative to non-US firms.

The presence of a Big-5 auditor and a listed ADR also appear to improve textual reporting. Firms with Big-5 auditors provide more extensive disclosure with less boilerplate and greater similarity with non-US firms, consistent with Hypotheses 1 a-c. Cross listing significantly increases the length of the annual report and increases similarity of reporting with both US and non-US firms, consistent with Hypotheses 1 a-d, although it is not significantly correlated with boilerplate. Results in Panel B with firm fixed effects are consistent with those in Panel A for IFRS and US GAAP, but are weaker with respect to ADR and Big-5 auditor because of their limited variability in the dataset.

The one characteristic that is mixed is Fog. Fog is positively related to IFRS and ADR. The coefficients on US GAAP and Big-5 auditor are positive but insignificant. While we do not have a directional prediction for Fog, it appears that Fog is higher under nonlocal accounting standards and greater oversight, perhaps reflecting greater disclosure of complex topics.[19]

Because the various text attributes are not independent and for parsimony going forward, we combine the textual analysis variables into factors using exploratory factor analysis. Results

---

[19] These and all other regressions are robust to additional controls for percent of closely held shares and the number of exchanges on which the firm is listed. However, inclusion of these variables significantly reduces sample size.

of the factor analysis are reported in Table 5. There appear to be two primary factors in the

textual variables. Factor 1 is strongly influenced by the quantity and comparability of disclosure.

It is higher for annual reports that are longer and are characterized by greater similarity with both

US and non-US firms. If we expect longer annual reports that are more comparable to be more

informative, we would expect Factor 1 to be "good" in the sense that it is associated with more

informative reporting. Factor 2, which is weaker, is higher for annual reports with more

boilerplate and Fog, suggesting that it captures less informative reporting.[20]

Table 4 also reports regression results including these two factors. Consistent with the

results from the previous analysis, Factor 1 tends to be larger for firms following IFRS or US

GAAP and for firms listing on US exchanges or employing Big-5 auditors. Results for Factor 2

are weaker, but largely consistent with the expectation that IFRS and US GAAP improve the

quality of reporting. The incremental effect of ADR in Panel B is positive (although only

marginally significant) perhaps reflecting the increased regulatory disclosure requirements

associated with cross listing.

VIII.   Economic Consequences

Determinants of textual attributes are potentially of innate interest because regulators and

standard setters have discussed the importance of constructs such as comparability, boilerplate,

and quantity of disclosure. However it is also important to assess the association between textual

attributes and potential economic outcomes one would expect to be correlated with improved

information content. This is particularly an issue in this context because there is little extant

---

[20] Comparability with non-US firms and Fog affect both factors, although non-US comparability is more strongly associated with Factor 1 while Fog is more strongly associated Factor 2. The presence of Fog and comparability with non-US firms in both factors suggests that they can enhance or reduce information depending on the context.

research linking these types of textual variables to potential information-based outcomes and several of our textual variables have not been used previously in the literature.

We investigate three types of outcomes: liquidity, analyst following, and institutional ownership. We use two measures of liquidity, zero return days and bid-ask spreads. Zero return days is the percent of days during the year on which the stock price did not change and is predicated on the notion that days with zero returns are likely days on which no significant trading occurred. While it runs the risk of misclassifying some days, it provides the largest potential sample because it does not require bid-ask spreads or volume data and is less sensitive to measurement variation. Bid-ask spread is the difference between the bid and ask price divided by the average of the two. It captures the explicit transactions cost of trading but is available for a smaller proportion of firms and may not be measured consistently across exchanges. Analyst following reflects the number of IBES analysts providing earnings forecasts. Institutional ownership is measured using institutional holdings data from the Thomson Reuter's International Mutual Fund (TIMF) database and captures the percent of shares held by mutual funds.

Results are presented in Table 6 for the two primary factors and we discuss results for the components in the text. Coefficient estimates for the control variables are largely consistent with expectations. In particular, liquidity, institutional ownership, and analyst following are higher for large firms, firms with positive profits, and firms with higher book/market ratios.

In terms of our primary relations of interest, results are consistent with Hypotheses 2 and 3, and with the correlations in Table 3. Factor 1, which captures the quantity and comparability of disclosure, is significantly negatively correlated with zero return days and positively correlated with analyst following and institutional ownership, consistent with Hypothesis 2. In terms of the underlying components, the results for the first factor are driven by the length of the

22

annual report and similarity with US firms. Factor 2 is positively correlated with both illiquidity measures and negatively correlated with analyst following and mutual fund ownership, consistent with Hypothesis 3. The results for the second factor are driven primarily by boilerplate, as well as Fog.

Overall, the results suggest that our textual attributes are correlated with economic outcomes in predictable ways. We are hesitant to draw causal inference given the difficulty in assessing whether it was the annual report disclosure per se that caused the information effects. Nevertheless, the results suggest that, as expected, the quantity of disclosure and similarity with US reporting are associated with improved information environments, while boilerplate and Fog are associated with less information.

IX. Mandatory IFRS Adoption and Disclosure Attributes

The preceding analyses suggest a general relation between textual attributes, firm characteristics, and economic outcomes. As noted in our discussion of Hypothesis 4.a, it is also informative to focus specifically on changes around mandatory IFRS adoption for at least two reasons. First, wide-spread adoption of IFRS represents one of the most significant innovations in the history of accounting and little is known about changes in associated textual disclosure. Second, because mandatory IFRS adoption was largely exogenous from the firm's perspective, examining changes in reporting around IFRS adoption allows us to better identify disclosure effects associated with IFRS and validate our textual measures.

In Table 7, we estimate the effect of mandatory IFRS adoption on disclosure using a difference-in-differences approach where mandatory adoption firms comprise our treatment

group and non-adopting firms our control group.[21] Because the regression includes both firms

that switched to IFRS as well as those that did not, we are able to measure changes in reporting

for IFRS adopters controlling for changes which occurred for non-adopting firms. Results for

the control variables are generally consistent with those in Table 4.

Our primary variable of interest is the interaction between the indicator for the post-

adoption period and the indicator for mandatory IFRS adopters, POST x MANDATORY, which

captures the difference in disclosure between firms affected by mandatory IFRS adoption relative

to the control sample. Because they are collinear with the firm and year fixed effects, POST and

MANDATORY are omitted from the regression.

Consistent with Hypothesis 4.a, IFRS adoption is associated with an increase in the

length of annual reports relative to the control sample, likely reflecting the increased disclosure

associated with IFRS adoption. In addition, IFRS adoption is associated with a lower percentage

of boilerplate relative to the control sample.

The coefficient on POST x MANDATORY in the XUS_COMP regression in column 3 is

significantly positive, suggesting that IFRS adoption is associated with increased comparability

of textual disclosure across non-US firms. To examine the comparability effect in more detail,

column 4 reports results for comparability among IFRS-adoption groups (i.e., adopters are

compared only with adopters and non-adopters are compared with non-adopters). While we

expect IFRS to increase comparability in general, the result should be particularly pronounced

relative to IFRS firms. Comparing column 3 with column 4, the effect of IFRS on comparability

was substantially stronger among IFRS firms (IFRS_COMP) than among all non-US firms in

general (XUS_COMP). While prior findings on accounting comparability around IFRS adoption

---

[21] We include 2005 in the post-IFRS sample since IFRS was mandatory in that year. However, results are similar if
we exclude 2005 from the analysis.

are mixed, our results suggest that the effects on textual comparability were quite pronounced, particularly among IFRS firms.

IFRS adoption is also associated with improved comparability with US firms (US_COMP). The results for textual comparability are consistent with the findings in prior literature that IFRS increased comparability in the measurement of net income and shareholders' equity between IFRS and US firms. The relation between IFRS adoption and Fog is insignificant. In terms of the two factors, IFRS adoption is associated with a significant increase in Factor 1, driven primarily by the increase in length and comparability around IFRS adoption. Factor 2 is significantly negative, reflecting primarily the decrease in boilerplate associated with IFRS adoption.

Overall, the results from the IFRS analysis suggest striking changes in textual characteristics around IFRS adoption, consistent with Hypothesis 4.a. While it is dangerous to draw strong conclusions, it seems plausible based on our results that the effects of IFRS on textual disclosure were at least as important as the effects on accounting measurement. Further, the IFRS results help to mitigate concerns about endogeneity and omitted correlated variables.

In addition, we conducted untabulated analyses splitting the sample and find systematic differences in the effects of IFRS across various country- and firm-level partitions. Previous research suggests that differences in reporting philosophies and regulation, in particular code versus common law, affect accounting practice. The direction of the effect is difficult to predict because the philosophy underpinning IFRS is arguably more consistent with common law reporting and enforcement of IFRS adoption is likely stronger in common law countries. On the other hand, local GAAP in code law countries is arguably less similar to IFRS, causing these firms to potentially make larger adjustments to comply. Similarly, firms with Big-5 auditors

may apply IFRS more conscientiously, increasing the effects of IFRS, but they may also have had higher quality reporting prior to IFRS adoption. To the extent that non-English language countries translated their annual reports to English (as opposed to authoring the originals in English), IFRS adoption may have affected the English reports and translation process for firms in non-English countries differently relative to reports that originated in English.

In general, the effects of IFRS adoption appear to have improved Factor 1 components more for code law countries and non-English countries, consistent with the notion that these countries likely provided less, and more idiosyncratic, disclosure prior to IFRS adoption. In terms of Factor 2, common law and English language countries experienced the greatest improvements, primarily driven by reductions in boilerplate. Overall Big-5 auditors were associated with greater improvements in Factor 1, suggesting that better oversight was associated with greater improvements in positive disclosure attributes around IFRS adoption.

X.   Changes in Annual Report Topics around IFRS Adoption

To provide further insight into what specific aspects of reporting changed most around IFRS adoption, we use a textual methodology developed in the natural language processing literature called Latent Dirichlet Allocation (LDA). As discussed in more detail in the Appendix, LDA uses the probability of words co-occurring within documents to identify sets of topics and their associated words. It can be thought of as the textual equivalent of factor analysis, where the model produces topics instead of factors.

We examine which topics experienced large increases in prominence after IFRS adoption by estimating the effect of IFRS adoption on the relative proportion of topics in firm annual reports, measured by the topic loading. We estimate separate OLS regressions with firm and year

fixed effects for each topic, where the topic loading is the dependent variable and the main independent variable of interest is the POST x MANDATORY interaction. The coefficient on this variable can be interpreted as the effect of IFRS adoption on the extent to which firms discussed a particular topic in their annual reports. We control for length to ensure that topic loadings were not altered as a mechanical function of changes in document length.[22] We expect topics related to standards which required substantial additional disclosure under IFRS relative to local GAAP to show the greatest increases in length after adoption. For example, required disclosures on financial instruments and derivatives (IFRS 7), employee compensation (IAS19 and IFRS2), and more detailed operating information (IFRS 3), among others, are particularly extensive under IFRS and thus would potentially show substantial increases after adoption.

Table 8 reports the coefficients and t-statistics for the POST x MANDATORY variable for those topics which showed the largest increases around IFRS adoption, as well as the twenty most prominent words within each topic. After IFRS adoption, firms appeared to substantially increase the amount of disclosure dedicated to financial instruments and derivatives (Topic 2), executive compensation (Topics 5 and 6), and employee stock options (Topic 1). Some of the keywords for the topics are more difficult to interpret, but reading reports for occurrence of the key terms suggests that Topic 3 relates to the description of a firm's accounting policies, while the other topics appear to cover financial performance of, for example, subsidiaries (Topic 9) and investments (Topic 8). While the results are descriptive, they suggest that the increases in disclosure length around mandatory IFRS adoption appear to be associated with specific disclosures that one might expect to increase under IFRS.

---

[22] The topics we identify as exhibiting the largest increases around IFRS adoption are robust to including additional controls or omitting them altogether. Although we report the coefficients and t-statistics from the mandatory adopter specification, results are consistent including voluntary adopters.

XI.   Instrumental Variables Analysis of Economic Outcomes

Our discussion of the effects of our textual characteristics on market outcomes in Table 6 was caveated by the fact that the results could be influenced by correlated omitted variables. In this section, we utilize mandatory IFRS adoption and attributes of peer firms in an instrumental variables analysis to more robustly examine the relation between firm disclosure and liquidity, analyst following, and institutional ownership.[23]

In order to perform the instrumental variables (IV) analysis, we need to identify at least as many exogenous instruments as endogenous regressors. A good instrument exhibits a high correlation with the endogenous variable to instrument but is uncorrelated with the error term. Factors 1 and 2 in Table 6 are both potentially endogenous, so we need at least two instruments in order to identify the IV specification. Table 4 indicates that various other attributes such as auditor and ADR listing are associated with textual characteristics, but these variables also reflect firm choices. However, mandatory IFRS adoption was relatively exogenous at the firm level. Additionally, firm-level disclosure is likely to be correlated with the disclosure of industry peers because of similarities in industry accounting and disclosure requirements, but peer disclosure is also largely exogenous from the perspective of the firm. We therefore use mandatory IFRS adoption and the attributes of peer firms in the same country-industry-year as instrumental variables to identify the outcomes associated with disclosure.

Table 9 Panel A reports results of regressions of FACTOR1 and FACTOR2 on our instruments, where IFRS is an indicator variable coded 1 if the firm has mandatorily adopted IFRS, and MEAN_PEER_FACTOR1 and MEAN_PEER_FACTOR2 are the mean levels of FACTOR1 and FACTOR2 for all other firms in the same country-industry-year. MEAN_PEER_ADR and MEAN_PEER_BIG5 are the percent of a firm's industry peers that

---

[23] We thank the referee for suggesting this approach.

have an ADR or Big 5 auditor, respectively.[24] The results in the table indicate that IFRS,

MEAN_PEER_FACTOR1, and MEAN_PEER_FACTOR2 have predictable, statistically

significant effects on FACTOR1 and FACTOR2. The other two instruments do not appear to be

particularly strong. Overall, our instruments for FACTOR1 and FACTOR2 are powerful enough

to be able to predict textual characteristics, as indicated by the significant F-statistics.[25]

Table 9 Panel B reports the results of our IV analysis using the instruments in Table 9

Panel A for FACTOR1 and FACTOR2. The results are very similar to Table 6 and indicate that

textual firm disclosure has a significant association with liquidity, analyst following, and

institutional ownership.[26] A drawback to using mandatory IFRS adoption as an instrument for

firm-level disclosure is that doing so excludes firms that voluntarily adopted IFRS. Also, we

document in Table 10 that IFRS has a direct effect on economic outcomes which means that

omitting it in the second stage can introduce omitted variables bias. Therefore, in Table 9 Panel

C we estimate an instrumental variables specification using only peer variables as instruments.

Consistent with Table 6, we again find significant associations between textual disclosure and

liquidity, analyst following, and institutional ownership.[27] Subject to the caveat that this

approach may not entirely resolve the endogeneity issue because the validity of our findings

depends on the appropriateness of our instruments, these results provide some comfort that firm-

level choices do not drive our empirical results.

---

[24] Industry is defined at the 2-digit ICB level. Country-industry-years with fewer than 3 total firms are excluded. All firms in this test are either mandatory adopters or non-adopters.

[25] Table 9, Panel A does not report the actual first stage regressions, which vary for each outcome variable based on data requirements and which include (following standard practice) the other control variables from the final specification which are not exogenous. This panel is for illustrative purposes to demonstrate instrument strength.

[26] IFRS is not included in the second stage because the exogenous instruments can only be included in the first stage. All other control variables from Table 6 are included. We find similar results if we use only mandatory IFRS adoption as an instrument for FACTOR1 and FACTOR2. However, the outcomes regressions can only be estimated including one of the two factors at a time in order for the equation to be identified.

[27] In robustness tests, we estimate the specifications in Panels B and C of Table 9 with additional controls for the average level of all control and outcome variables for the industry peers to mitigate the possibility that our results could be driven by the effect of peer incentives (which could be correlated with firm-level incentives and disclosure) or by peer economic outcomes (which could be correlated with firm-level outcomes). Inferences are unchanged.

XII.   Mandatory IFRS Adoption and Economic Outcomes

In our final analysis, we examine the association between cross-sectional differences in the effect of mandatory IFRS adoption and our economic outcome variables. We identify firms whose textual disclosure appears to have significantly improved around IFRS adoption and compare their outcomes to firms which show less improvement. Our prediction from Hypothesis 4.b is that firms which experienced the greatest increase in "desirable attributes" around IFRS (disclosure length and comparability based on Factor 1 above) experienced greater increases in liquidity, analyst following, and institutional investment, while those that experienced increases in "undesirable attributes" (boilerplate and Fog based on Factor 2 above) experienced reduced benefits. This analysis allows us to investigate the extent to which the benefits of IFRS adoption varied based on its impact on firms' textual reporting.

To identify high and low quality IFRS implementers, we examine changes in disclosure in the year of IFRS implementation. Firms that experienced increases in Factor1 (length and comparability) in the top quintile of firms in 2005 are identified as "high benefit implementers" (HBen=1), while firms that experienced large increases in Factor 2 (boilerplate and Fog) are identified as "low benefit implementers" (LBen=1).[28] This approach is similar to Daske et al. (2013) who identify "serious" adopters of IFRS by isolating those which had large changes in accounting accruals, reporting incentives, or analyst following in the year of IFRS adoption.

Table 10 Panels A and B report the results of these tests. The dependent variables are the four outcome variables introduced in Table 6—liquidity, analyst following, and institutional ownership. The independent variables of interest are the variables relating to IFRS adoption. As in Table 7, we include POST x MANDATORY to capture the difference-in-differences effect of

---

[28] Although we classify firms based on their disclosure implementation in 2005, results are similar when we discard the actual year of or after adoption, or both, suggesting that the effects we document are persistent.

mandatory IFRS adoption, but here we are measuring the effect on our economic outcomes instead of textual attributes. To the extent that mandatory IFRS adoption improved financial reporting in general, we expect to observe an increase in liquidity, analyst following, and institutional ownership around IFRS adoption. In addition, consistent with Hypothesis 4.b, we expect the benefits of mandatory adoption to vary based on its effect on the textual variables. We predict that high benefit implementers (HBen=1) will receive greater benefits of IFRS adoption than low benefit implementers (LBen=1). We test that prediction by interacting the POST x MANDATORY variable with our high and low benefit implementation indicators (HBen and LBen). We expect the interaction POST x MANDATORY x HBen to be positively associated with liquidity, analyst following, and institutional ownership, and the interaction POST x MANDATORY x LBen to be negatively associated with liquidity, analyst following, and institutional ownership.[29]

The results in Table 10, Panel A are consistent with the predictions in Hypothesis 4.b. The positive (negative) coefficient on POST x MANDATORY for analyst following (illiquidity) suggests that mandatory IFRS adoption had a positive effect on liquidity and analyst following overall. More importantly, the interactions with HBen and LBen suggest that firms with greater increases in "good" disclosure attributes experienced relatively greater benefits to liquidity, analyst following, and institutional ownership, while firms with increases in "bad" disclosure attributes had reduced benefits of adoption.[30]

---

[29] The POST x MANDATORY x HBen (LBen) interactions capture the incremental effects for firms with the greatest changes in textual attributes relative to the typical IFRS-adopting firm. We omit the main effects for POST, MANDATORY, HBen and LBen because they are redundant given the inclusion of firm and year fixed effects.
[30] Based on the point estimates, low quality implementers experienced essentially no benefit to analyst following and a negative change in institutional ownership around IFRS adoption. The notion that beneficial effects of IFRS were limited to subsets of firms based on implementation is consistent with the results in Daske et al (2013).

A potential issue with the preceding approach, which is shared with other research such as Daske et al. (2013), is the fact that, while IFRS was largely exogenous from the standpoint of the firm, variation in disclosure changes around IFRS adoption may not only reflect exogenous variation in the effects of new disclosure requirements but also firm-level implementation choices. To abstract from firm-level choices, we use the fact that IFRS requirements likely differed systematically in terms of how they affected the disclosure in particular industries to specify an instrumental variables analysis that focuses on IFRS effects at the industry level. For example, the impact of standards such as financial instruments and derivatives (IFRS 7), employee benefits and pensions (IAS19), share-based payments (IFRS 2), provisions and contingencies (IAS 37), property, plant and equipment (IAS 16), intangibles (IAS 38), asset impairments (IAS 36), and leases (IAS 17) likely varied based on industry.

We exploit this potential variation in the effects of IFRS in the instrumental variables analysis in Table 10, Panel B. Our first-stage instruments are the indicator variables for 3-digit ICB industry, as well as their interactions with the POST x MANDATORY variable.[31] Although the results are weaker in Panel B, we find consistent evidence that differences in disclosure around IFRS adoption were associated with firm outcomes. Firms in industries for which increases around IFRS adoption were greater for Factor 1 (length and comparability) experienced significantly higher increases in liquidity as measured by zero return days as well as analyst following, while firms in industries for which increases around IFRS adoption were greater for Factor 2 (boilerplate and Fog) experienced significant reductions in the effect of IFRS on

---

[31] The interactions are included as instruments to avoid the "forbidden regression" problem discussed in Wooldridge (2002) and demonstrated in an accounting context in Beaver et al. (2012), and are especially important in this analysis because the industry dummies themselves are perfectly collinear with the firm fixed effects and are thus omitted. The F-statistics of the excluded instruments have p-values below the 1% level in all first-stage regressions.

liquidity as measured by zero return days and analyst following.[32] Subject to the caveat that the IV results in Panel B may not entirely address endogeneity and could be biased if our instruments are sufficiently weak, these results provide evidence of variation in the benefits of IFRS adoption across firms in addition to the overall IFRS effects considered in our earlier analyses.

XIII. Conclusions

There is a substantial existing literature on determinants and consequences of accounting measurement, particularly with respect to net income and shareholders' equity, for broad samples of global firms and around major events such as IFRS adoption. While accounting measurement is clearly important, the associated textual disclosure is likely to be at least as important, yet it has attracted much less attention due, at least in part, to the difficulty in gathering large samples of non-US annual reports and in quantifying textual data. Our goal in this paper is to provide some of the first evidence on determinants and outcomes of annual report textual characteristics for a broad sample of non-US firms over time.

We focus on textual characteristics—quantity of disclosure, boilerplate, Fog, and comparability with US and non-US peer firms—that are likely to be of interest to regulators, investors, and other users of annual report information. Our results suggest that annual report textual disclosure quality appears to be higher (more disclosure, less boilerplate, and greater comparability with non-US and US peers) in contexts with more stringent accounting standards, stronger oversight, and greater demand for information (e.g., IFRS, US GAAP, cross listing, and

---

[32] Similar to Table 9, in robustness checks we include control variables in the second stage for average industry-year values of both the control variables and the outcome variables to account for industry-level incentives that could be correlated with firm-level incentives and to remove any direct effects of industry-level outcome variables on firm-level outcomes. Inferences are unchanged.

Big-5 auditing). Further, our textual attributes are correlated with economic outcomes such as liquidity, analyst following, and mutual fund ownership. This finding is robust to an instrumental variables analysis using mandatory IFRS adoption and peer attributes as instruments. Finally, differences in the effects of IFRS adoption on textual disclosure are associated with variation in market outcomes; firms with the greatest increases in beneficial disclosure (length and comparability) benefit more from IFRS while those with increases in poor disclosure attributes (boilerplate and Fog) experienced smaller benefits.

While admittedly somewhat exploratory, we believe our results are likely to be of interest to regulators, standard setters, investors, and other researchers going forward. There is clearly scope for additional research focusing on more specific aspects of textual disclosure and developing more sophisticated measures. Our conclusions are subject to important caveats. While our measures appear to behave in a manner consistent with expectations, they are exploratory. Further, it is difficult to infer causality, even using shocks such as mandatory IFRS adoption. Finally, our sample is limited to English language reports and thus cannot speak to the effects of non-English disclosure. Overall, though, we view our analysis to be a potentially useful first step in a developing important research paradigm.

References

Ahmed, A., M. Neel and D. Wang, 2013. Does Mandatory Adoption of IFRS Improve Accounting Quality? Preliminary Evidence. *Contemporary Accounting Research* 30, 1344-1372.

Alford, A., J. Jones, R. Leftwich and M. Zmijewski, 1993. The Relative Information Content of Accounting Disclosures in Different Countries. *Journal of Accounting Research* 31, 183-223.

Amihud, Y., 2002. Illiquidity and Stock Returns: Cross-section and Time-series Effects. *Journal of Financial Markets*, 5: 31–56

Armstrong, C., M. Barth, A. Jagolinzer and E. Riedl. 2010. Market Reaction to the Adoption of IFRS in Europe. *The Accounting Review* 85: 31-61.

Ball, C., G. Hoberg, V. Maksimovic, 2013. Disclosure Informativeness and the Tradeoff Hypothesis: A Text-Based Approach. Working Paper.

Barrett, R., C. Lyon and J. Malcolm, 2006. Plagiarism Is Easy, But Also Easy To Detect. Plagiary: Cross-Disciplinary Studies in Plagiarism, Fabrication, and Falsification, 57-65.

Barth, M., W. Landsman and M. Lang, 2008. International Accounting Standards and Accounting Quality. *Journal of Accounting Research* 46, 467-498.

Barth, M., W. Landsman, M. Lang and C. Williams, 2012. Are IFRS-based and U.S. GAAP-based accounting amounts comparable? *Journal of Accounting and Economics* 54 (1): 68-93.

Beaver, W., W. Landsman, E. Owens, 2012. Asymmetry in Earnings Timeliness and Persistence: A Simultaneous Equations Approach. *Review of Accounting Studies*, 18: 781-806.

Bischof, J. and H. Daske, 2013. Mandatory Disclosure, Voluntary Disclosure, and Stock Market Liquidity: Evidence from the EU Bank Stress Tests. *Journal of Accounting Research*, 51: 997-1029.

Blankespoor, E., 2014. The Impact of Information Processing Costs on Firm Disclosure Choice: Evidence from the XBRL Mandate. Working Paper.

Blei, D., A. Ng, M. Jordan, 2003. Latent Dirichlet Allocation. *The Journal of Machine Learning Research*, 3: 993-1022.

Brochet, F., P. Naranjo and G. Yu, 2013, Capital Market Consequences of Linguistic Complexity in Conference Calls of Non-U.S. Firms. Working Paper.

Brown, S. and J. Tucker. 2011. Large-Sample Evidence on Firms' Year-over-Year MD&A Modifications. *Journal of Accounting Research*, 49: 309–346.

Bushee, B., I. Gao and D. Taylor, 2014, Linguistic Complexity in Firm Disclosures: Obfuscation or Information? Working Paper.

Cascino, S. and J. Gassen, 2012. Comparability Effects of IFRS Adoption. Working Paper.

Christensen, H. L. Hail and C. Leuz, 2013. Mandatory IFRS Reporting and Changes in Enforcement. *Journal of Accounting and Economics,* 56: 147-177.

Daske, H., L. Hai, C. Leuz and R. Verdi, 2013. Adopting a Label: Heterogeneity in the Economic Consequences Around IAS/IFRS Adoptions. *Journal of Accounting Research,* 51: 495-547.

De Franco, G., S.P. Kothari and R. Verdi, 2011. The Benefits of Financial Statement Comparability. *Journal of Accounting Research,* 49: 895–931.

Ernstberger, J. and M. Grüning, 2013, How do Firm- and Country-level Governance Mechanisms Affect Firms' Disclosure? *Journal of Accounting and Public Policy,* 32 (2):5-67

Financial Accounting Standards Board, 2010, Conceptual Framework for Financial Reporting, Statement of Financial Accounting Concepts No. 8, Norwalk, CT.

Financial Accounting Standards Board, 2012, Disclosure Framework. Discussion Paper, Norwalk, CT.

Hoberg, G. and C. Lewis, 2014. Do Fraudulent Firms Strategically Manage Qualitative Disclosure? Working Paper.

Hoberg, G. and G. Phillips, 2010. Product Market Synergies and Competition in Mergers and Acquisitions: A Text-based Analysis. *Review of Financial Studies*; 23: 3773-811.

Hoogervost, H., 2013. Breaking the Boilerplate, Remarks to the IFRS Conference, Amsterdam, June 27.

Huang, A., R. Lehavy, A. Zang, R. Zheng, 2014. A Thematic Analysis of Analyst Information Discovery and Information Interpretation Roles. Working Paper.

Jeanjean, T., C. Lesage and H. Stolowy, 2010. Why Do You Speak English (in your annual report)? *The International Journal of Accounting* 45 (2): 200-223.

Jeanjean, T., H. Stolowy, M. Erkens and T. Yohn. 2013. Evidence on the Impact of Adopting English as an External Reporting Language on Foreign Investment, Liquidity, and Analyst Following. Working Paper.

Land, J., and M. Lang. 2002. Empirical Evidence on the Evolution of International Earnings. *The Accounting Review* 77: 115-134.

Lang, M., K. Lins and M. Maffett, 2012. Transparency, Liquidity, and Valuation: International Evidence on When Transparency Matters Most. *Journal of Accounting Research*, 50: 729–774.

Lang, M. and R. Lundholm. 1996. Corporate Disclosure Policy and Analyst Behavior. *The Accounting Review* 71, 467-492.

Lang, M. and Maffett, M. 2011. Transparency and Liquidity Uncertainty in Crisis Periods. *Journal of Accounting and Economics* 52, 101–25.

Lehavy, R., F. Li and K. Merkley. 2011. The Effect of Annual Report Readability on Analyst Following and the Properties of Their Forecasts. *The Accounting Review* 86: 0187-1115.

Li, F., 2008. Annual Report Readability, Current Earnings, and Earnings Persistence. *Journal of Accounting and Economics* 43:221-247.

Li, F., 2010. Textual Analysis of Corporate Disclosures: A Survey of the Literature. *Journal of Accounting Literature*. 29: 143-165.

Li, H., 2014. Repetitive Disclosures in the MD&A. Working Paper.

Lundolm, R., R. Rogo and J. Zhang, 2013. Restoring the Tower of Babel: How Foreign Firms Communicate with US Investors. Working Paper.

Maffett, M., 2012. Financial Reporting Opacity and Informed Trading by International Institutional Investors, *Journal of Accounting and Economics* 54 (2–3): 201-220.

Miller, B. 2010. The Effects of Reporting Complexity on Small and Large Investor Trading, *The Accounting Review*, 85, 2107-2143

McCallum, A., 2002. MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu.

McMullin, J., 2013. Can I Borrow Your Footnotes? Footnote Similarity and Accounting Comparability. Working Paper.

Nelson, K and A. Prichard, 2007. Litigation Risk and Voluntary Disclosure: The Use of Meaningful Cautionary Language. Working Paper.

Peterson, K., R. Schmardebeck and T. Wilks, 2012. Accounting Comparability and Earnings Attributes. Working Paper.

Securities and Exchange Commission, 1998, "A Plain English Handbook: How to Create Clear SEC Disclosure Documents. Washington, DC.

Wooldridge, J., 2002. *Econometric Analysis of Cross Section and Panel Data.* Cambridge: The MIT Press.

Wysocki, P., 2008. Assessing earnings and accruals quality: US and international evidence. Working Paper.

Yip, R., and D. Young, 2012. Does Mandatory IFRS Adoption Improve Information Comparability? *The Accounting Review*: September 2012, Vol. 87, No. 5, pp. 1767-1789.

Variable Descriptions

Textual Attributes

> BOILERPLATE = the percent of words in the annual report that are in boilerplate
> sentences.

> FACTOR1 = the first factor produced in a factor analysis of the main textual
> characteristics, driven mostly by LN_NBWORDS, XUS_COMP, and US_COMP.

> FACTOR2 = the second factor produced in a factor analysis of the main textual
> characteristics, driven mostly by FOG and BOILERPLATE.

> FOG = the Gunning Fog index calculated as (words_per_sentence +
> percent_complex_words)*0.4.

> IFRS_COMP = a version of the XUS_COMP measure that is only calculated within
> IFRS adoption peer groups. That is, non-adopters are only compared with non-
> adopters, and adopters are only compared with other adopters.

> HBen = an indicator variable equal to one if the change in FACTOR1 in the year of IFRS
> adoption was in the top quintile for all firms.

> LBen = an indicator variable equal to one if the change in FACTOR2 in the year of IFRS
> adoption was in the top quintile for all firms.

> LN_NBWORDS = the natural log of the total number of non-boilerplate words in firm i's
> annual report.

> LN_WORDS = the natural log of the total number of words in firm i's annual report.

> MEAN_PEER_FACTOR1 = the average level of the variable FACTOR1 for firms in the
> same country-industry-year (where industry is the 2-digit ICB code), excluding
> the current firm.

> MEAN_PEER_FACTOR2 = the average level of the variable FACTOR2 for firms in the
> same country-industry-year (where industry is the 2-digit ICB code), excluding
> the current firm.

> NUM_WORDS(x) = the total number of words in the annual report (to the xth power).

> TOPIC_N = the topic loading of the Nth LDA topic in a given topic (interpretable as the
> weight of Topic N in the document).

> US_COMP = the median aggregate cosine similarity of the annual report of a firm with a
> matched sample of US firms in the same industry after stemming and stopwording

procedures and after deleting all words not found in the Loughran McDonald English business dictionary.

XUS_COMP = the median aggregate cosine similarity of the annual report of a firm with a matched sample of non-US firms outside of the firm's home country that are in the same industry after stemming and stopwording procedures and after deleting all words not found in the Loughran McDonald English business dictionary.

Outcome Variables

ANALYST = the number of unique analysts issuing forecasts for firm i's earnings.

LN_BIDASK = the median bid ask spread over the fiscal year, where the bid ask spread is defined as (ask–bid)/((ask+bid)/2). We take the natural log to reduce skewness.

INST_OWN = the percent of the firm's total common stock outstanding that is currently held by institutional investors, constructed using data from the Thomson Reuter's International Mutual Fund (TIMF) database. See Maffett (2012) for more information on this database.

ZERO_RETURN = the percent of the total trading days in the year where the firm had a stock return of zero.

Independent Variables

ADR = an indicator variable coded 1 if the firm has an American Depository Receipt.

AGE = age of a firm in years, approximated using its date of initial coverage in Datastream, following Maffett (2012).

BIG5 = an indicator variable coded 1 if the firm has a Big-5 auditor.

BM_RATIO = book-to-market ratio, using book value of common equity divided by market value of common equity.

EARN_SURPRISE = change in earnings per common share, scaled by price of common shares at the end of the prior year.

IFRS = an indicator variable coded 1 if the firm uses International Financial Reporting Standards.

LEVERAGE = total debt (short-term + long-term) divided by total assets.

LN_MVE = the log of the firm's market value of equity (thousands).

MANDATORY = an indicator variable coded 1 if a firm mandatorily adopted IFRS in 2005.

MEAN_PEER_ADR = the percent of firms in the same country-industry-year (where industry is the 2-digit ICB code) that have an ADR, excluding the current firm.

MEAN_PEER_BIG5 = the percent of firms in the same country-industry-year (where industry is the 2-digit ICB code) that have a Big 5 auditor, excluding the current firm.

POST = an indicator variable coded 1 if the fiscal year is 2005 or later.

ROA = return on assets. Net income before extraordinary items divided by total assets.

NI_LOSS = an indicator variable coded 1 if the firm reports a loss.

US_GAAP = an indicator variable coded 1 if the firm uses US Generally Accepted Accounting Principles.

(XUS/US/IFRS)_ROA_DIST = the median absolute difference between a firm's return on assets and that of the firm peers used to calculate its COMP, for both the US and non-US peers, as well as non-US peers within IFRS groups.

(XUS/US/IFRS)_SIZE_DIST = the median difference between a firm's size (lagged total assets) and that of the firm peers used to calculate its COMP, divided by lagged total assets. Calculated for both US and non-US peers, as well as non-US peers within IFRS groups.

Appendix

Data Treatment

We convert the PDF version of the annual reports to text files using the Xpdf and QPDF software programs. The conversion of PDF files to text can sometimes lead to garbled output and makes it difficult to identify tables. To deal with this, we first use the Lingua::EN::Sentence Perl module to break the text of each report into sentences. Then we remove all "sentences" that do not contain at least 50% alphabetic characters, similar to Li (2008) and Miller (2010), and delete sentences where more than 20% of the characters are not alphanumeric (usually because they contain foreign language characters or symbols added by a conversion error). We also exclude lines consisting of fewer than 50 alphabetic characters, for example page headings.

This procedure does not successfully delete all table labels because the PDF conversion process can separate the numbers and labels in tables into separate "sentences." This adds noise to the fog measure, which is not designed to analyze this type of content, but is appropriate to include in the other textual measures, such as comparability, under the assumption that the text in tables is widely read and includes relevant information for financial statement users.

We use the Lingua::EN::Fathom module to calculate the Fog score, total word count, and document word vectors for the remaining text. We further process these word vectors before using them to calculate comparability by removing all stop words such as "the" and "and" and by stemming the remaining words.[33] We require each of these word stems to be present in the Loughran and McDonald business word list (February 2013 version).[34] Restricting the words to this dictionary helps us avoid including proper nouns or words from foreign languages.

---

[33] We use the stop words list in the Lingua::Stopwords Perl module. Stemming is the process of removing grammatical endings from words in order to combine words with similar meanings. We stem words using the Lingua::Stem::En Perl module which uses the Porter stemming algorithm.

[34] For more information on the word list see http://www3.nd.edu/~mcdonald/Word_Lists.html.

The preceding procedures generally do a good job at editing and cleaning English documents but are not able to identify all foreign language documents or all documents with conversion errors. To exclude these, we delete annual reports that have fog scores less than 12 or greater than 30, those that do not have at least 50 sentences, those with fewer than 100 words, those with fewer than 30 unique word stems, and those where the ratio of the total number of words to the number of unique word stems is greater than 60. Next, we delete all observations where the number of words in the final document is above (below) the 99th (1st) percentile within the firm's country. Lastly, we exclude all firm-years where the firm's annual report appears to be an SEC filing in order to prevent US regulatory language being spuriously linked with determinants or outcomes of disclosure when they are really driven by SEC requirements.

In order to merge our annual reports with financial data, we match the company name associated with each report with company names in the Osiris and Datastream databases. Most names do not have a perfect match, so we standardize common business terms (such as "Incorporation" to Inc) and delete stop words. For any remaining names we fuzzy match using the Compged command in SAS. We accept matches with a score of 200 or less (for comparison, the score for "Gold Mine Inc" and "Gold Mines Inc" is 100). We hand checked remaining unmatched firms for countries which still had more than five percent of the names unmatched. Most of the remaining unmatched firms were delisted firms or subsidiaries of larger companies.

Comparability Measures

To construct the comparability measures, we first calculate the cosine similarity of a firm's annual report with each of its non-US and US peers, respectively, by taking the dot product of their document word vectors scaled by the product of their lengths:[35]

---

[35] The word vectors weight each word by the number of times it occurs in the document, but there are other weighting schemes. Peterson et al. (2012) construct vectors of 0's and 1's and Brown and Tucker (2011) use the tf-

$$(A1)\ Similarity_{AB} = \cos(\theta) = \frac{A \cdot B}{||A||\,||B||}$$

By scaling by the lengths of the vectors, cosine similarity abstracts from differences in documents driven solely by length.[36] Our final comparability measure is the median[37] cosine similarity of a firm's annual report with all of its peer firms, where non-US 3-digit ICB peers outside of the home country in the same industry-year are used to calculate XUS_COMP and US 3-digit ICB peers are used to calculate US_COMP:[38]

$$(A2)\ XUS\_COMP_{it} = Median\ (XUS\_Similarity_{ijt})$$

$$(A3)\ US\_COMP_{it} = Median\ (US\_Similarity_{ijt})$$

To ensure that regression results are not driven by the quality of our peer matches, we calculate four match quality measures that are the median distance between a firm's lagged total assets and its return on assets relative to its peer firms, for both US and non-US peers. We include these measures as controls in all regressions that contain comparability.

Boilerplate

We identify boilerplate disclosure by first counting all tetragrams contained in each document in our sample, where a tetragram is an ordered group of four words within a single sentence. We aggregate these counts by country and discard tetragrams that do not occur in at least 30% of the documents of a country or on average at least 5 times per document. Sentences

---

idf approach which gives lower weight to very common words. Our approach is suited to our setting because we are interested in the relative frequency of words and topics, not just their presence as in Peterson et al. (2012). We also mitigate the effect of extremely common words by discarding stop words. The tf-idf approach would be inappropriate if words common in a business setting (e.g. "income") were informative and relevant to comparability.

[36] Brown and Tucker (2011) show that cosine similarity can still be affected by document length because longer documents are more likely to contain any given word. They adjust cosine similarity for length by removing the variation explained by its first five moments. We include these moments in our regressions of the determinants of our comparability measures and BOILERPLATE to ensure that results are not driven by length.

[37] Results were similar when we instead used the mean.

[38] Yip and Young (2012) argue that a good measure of comparability will show that economically similar firms have higher comparability than economically dissimilar firms. In untabulated tests, we show that non-US and US peers in different 1-digit ICB industries have significantly lower comparability than our 3-digit peers.

43

that contain a tetragram that occurs in at least 60% of the documents in the firm's home country are flagged as unusually common and potentially boilerplate.[39]

Examination of these phrases indicates that this procedure can capture common innocuous phrases (such as "as a result of") and regulatory disclosure in addition to potential boilerplate phrases. Because we would like our measure of boilerplate to capture uninformative, formulaic disclosure that the firm includes in its annual report, we exclude these innocuous common phrases as well as regulatory phrases and phrases from the auditor's letter. We conjecture (and inspection of the data confirms) that these phrases can be identified by excluding the most common phrases across documents. Therefore, our final identification of boilerplate excludes sentences which include common tetragrams that appear in more than 80% of the sample documents within that country. Testing this approach in the familiar regulatory environment of the United States, we see that this simple rule removes phrases that are clearly linked with mandatory disclosure (for example the report on internal controls) as well as innocuous "grammatical" phrases.[40] We define our boilerplate measure, BOILERPLATE, as the percent of the total words in the annual report in boilerplate sentences.

It is possible that the remaining sentences may contain informative content. Therefore, in robustness tests we did not classify sentences as boilerplate if they contained potentially informative numbers (years, numbered accounting standards, item numbers, and numbered lists are not considered informative, following Blankespoor 2014). In another alternative measure, we only classified sentences as boilerplate if the combination of boilerplate phrases and

---

[39] Although trigrams (groups of three words) are commonly used to identify similar phrases across texts, Barrett et al. 2006 indicate that legal texts may require the use of tetragrams. While annual reports are not legal documents, they often contain many legal phrases and financial jargon. Tests using trigrams to identify boilerplate disclosure indicated that tetragrams fare better in this setting.

[40] In robustness, we separated regulatory and "innocuous" phrases using their average frequency within documents (regulatory phrases will occur in many documents but only once or twice per document). Allowing the "innocuous" phrases to be classified as boilerplate made no difference in the inferences of our study, but conceptually and empirically introduced noise into the measure.

(uninformative) stop words in the sentence comprised at least 60% of its length. These more restrictive measures of boilerplate are positively correlated with our main measure and did not affect inferences.

We are not the first to attempt to measure boilerplate disclosure. Nelson and Pritchard (2007) use word phrases to compare a firm's disclosure over time.[41] McMullin (2013) uses word phrases to identify parts of a firm's footnotes that are the same as those of matched peers to assess factors that drive similarity of disclosure across firms. Our measure differs because we capture disclosure that is redundant across many firms and therefore unlikely to be informative.

Latent Dirichlet Allocation (LDA)

LDA is an unsupervised machine-learning approach developed by Blei et al. (2003) to identify the topics contained in a large corpus of text.[42] Because LDA is unsupervised, it is replicable and free of researcher bias, but the topics can be difficult to interpret. The only parameter the researcher chooses is the number of topics to generate which requires some trial and error to avoid topics that are too broad or too narrow. We use the MALLET software developed by Andrew McCallum to generate the topics and topic loadings for our entire document collection, where the topic loading is a measure of how prominent a given topic is within an individual document and can be roughly interpreted as the proportion of the document comprising that topic. We generate 100 topics following Ball et al. (2013); topics were similar, though less interpretable, when we instead generated 150 topics.[43]

---

[41] We also calculated a measure of the similarity of a firm's disclosure over time, or disclosure "stickiness," using cosine similarity, similar to Brown and Tucker (2011). The resulting measure is positively correlated with our measure of boilerplate and our results are robust to its inclusion as a control.

[42] LDA has only recently been used in accounting and finance. Ball et al. (2013) use LDA to identify topics within the MD&A. Huang et al. (2014) examine the topics discussed in conference calls and analyst reports. Hoberg & Lewis (2014) examine changes in the content of a firm's MD&A around fraudulent behaviors that result in AAERs.

[43] We refer readers to the excellent discussion in Huang et al. (2014) for a more detailed explanation of the mechanics of LDA.

Table 1. Observations by Country Year

Country	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	Total
ARGENTINA	6	7	8	5	10	10	12	14	10	16	15	11	13	10	147
AUSTRALIA	117	177	297	580	548	662	686	664	1120	1236	1126	900	1128	157	9398
AUSTRIA	8	22	35	37	41	38	41	42	53	54	58	57	47	43	576
BELGIUM	11	32	33	62	57	56	49	54	70	67	72	58	59	63	743
BRAZIL	5	11	18	24	21	18	23	21	32	30	42	98	104	57	504
CANADA	298	433	539	592	698	658	616	758	935	1104	1065	1422	1388	1147	11653
CHILE	4	7	11	11	6	7	13	8	13	11	14	7	3	6	121
CHINA	.	15	37	55	85	95	106	121	110	114	129	153	194	232	1446
CZECH REPUBLIC	12	16	18	22	17	20	17	12	8	8	7	6	4	2	169
DENMARK	28	41	64	64	65	78	71	75	86	79	76	92	82	68	969
FINLAND	32	53	65	77	87	93	92	90	100	89	105	94	87	97	1161
FRANCE	37	75	99	110	138	159	157	171	141	166	177	190	182	132	1934
GERMANY	44	104	169	250	265	301	295	306	319	334	349	343	317	296	3692
GREECE	2	.	4	1	9	8	5	14	11	18	11	13	9	4	109
HONG KONG	40	57	130	171	183	181	178	175	181	179	165	177	183	147	2147
INDIA	8	24	50	60	238	272	358	408	719	804	827	1235	1234	52	6289
INDONESIA	2	6	31	35	29	29	49	60	83	119	133	139	172	225	1112
IRELAND	16	28	31	34	38	36	37	38	47	49	47	40	45	33	519
ISRAEL	5	8	21	32	34	39	42	42	57	50	47	39	31	20	467
ITALY	17	39	62	57	84	104	109	116	141	120	109	106	118	101	1283
JAPAN	296	447	514	552	438	472	594	609	446	405	357	337	219	14	5700
KOREA (SOUTH)	11	21	35	40	46	47	42	37	42	30	22	35	21	14	443
LUXEMBOURG	2	6	5	8	12	15	18	14	19	22	16	17	18	18	190
MEXICO	17	22	24	25	24	30	29	24	29	34	28	28	30	10	354
NETHERLANDS	14	42	75	74	96	95	91	89	90	63	94	93	80	78	1074
NEW ZEALAND	13	29	39	55	70	76	76	97	97	109	106	105	93	18	983
NORWAY	34	55	63	58	67	73	76	76	77	96	103	120	118	108	1124
PAKISTAN	5	5	10	11	20	28	36	60	81	89	72	89	67	17	590
PHILIPPINES	4	10	18	12	31	141	138	156	156	144	137	142	140	132	1361
POLAND	2	3	9	18	17	24	35	35	46	40	53	88	101	40	511
PORTUGAL	.	11	10	10	11	14	11	14	18	17	21	23	21	19	200
RUSSIA	.	.	2	6	7	8	12	11	19	20	25	35	34	33	212
SINGAPORE	23	68	126	281	325	349	337	330	449	443	443	471	456	337	4438
SOUTH AFRICA	18	42	92	127	158	157	173	186	181	171	203	230	230	80	2048
SPAIN	7	19	30	35	37	49	42	54	61	58	51	67	54	32	596
SRI LANKA	.	1	4	5	9	11	9	26	33	80	141	153	134	30	636
SWEDEN	55	90	113	121	138	146	156	151	163	170	148	138	163	123	1875
SWITZERLAND	36	72	111	127	144	141	151	154	163	164	163	169	155	159	1909
TAIWAN	1	9	12	22	46	55	38	44	59	56	56	70	31	29	528
THAILAND	8	15	30	230	272	280	294	310	381	395	402	415	266	403	3701
TURKEY	.	1	7	16	25	44	55	60	51	61	61	88	83	98	650
UNITED KINGDOM	238	595	770	841	935	1041	1113	1068	1338	1583	1554	1367	1089	514	14046
Total	1476	2718	3821	4953	5581	6160	6482	6794	8235	8897	8830	9460	9003	5198	87608

Table 2. Descriptive Statistics

Panel A: Primary Variable Descriptive Statistics

Variable	N	Mean	Median	Std	P25	P75
LN_WORDS	87,608	9.733	9.758	0.671	9.299	10.184
LN_NBWORDS	87,608	9.664	9.686	0.679	9.225	10.118
BOILERPLATE	85,911	7.045	7.052	3.649	4.136	9.635
XUS_COMP	84,997	0.682	0.690	0.069	0.643	0.732
US_COMP	86,060	0.632	0.636	0.059	0.598	0.671
FOG	85,793	19.520	19.561	1.731	18.270	20.805
ZERO_RETURN	82,253	27.974	17.375	27.676	5.118	45.783
LN_BIDASK	63,905	0.046	0.018	0.072	0.007	0.051
ANALYST	50,599	5.417	3.000	5.410	1.000	8.000
INST_OWN	52,161	10.736	6.702	11.859	1.432	16.011
LN_MVE	87,608	11.730	11.648	2.243	10.116	13.286
ROA	86,008	-0.029	0.025	0.236	-0.024	0.065
BIG5	87,608	0.499	0.000	0.500	0.000	1.000
IFRS	87,608	0.337	0.000	0.473	0.000	1.000
US_GAAP	87,608	0.012	0.000	0.109	0.000	0.000
ADR	87,608	0.008	0.000	0.090	0.000	0.000
BM_RATIO	86,195	0.984	0.723	1.022	0.389	1.220
LEVERAGE	86,125	0.209	0.168	0.201	0.019	0.337
EARN_SURPRISE	85,907	0.041	0.008	0.289	-0.031	0.057
NI_LOSS	87,608	0.309	0.000	0.462	0.000	1.000
AGE	87,608	13.129	11.157	9.352	5.676	17.829

Panel B: Primary Variable Descriptive Statistics for the Datastream Population

Variable	N	Mean	Median	Std	P25	P75
ZERO_RETURN	250,828	27.649	15.873	27.563	5.976	44.534
LN_BIDASK	190,983	0.043	0.016	0.071	0.006	0.047
ANALYST	119,563	4.859	2.000	5.392	1.000	7.000
INST_OWN	159,114	7.167	2.890	9.833	0.395	10.083
LN_MVE	266,295	11.409	11.266	2.218	9.910	12.786
ROA	260,972	-0.023	0.020	0.211	-0.017	0.057
BIG5	266,295	0.367	0.000	0.482	0.000	1.000
IFRS	266,295	0.216	0.000	0.412	0.000	0.000
US_GAAP	266,295	0.018	0.000	0.134	0.000	0.000
ADR	266,295	0.020	0.000	0.140	0.000	0.000
BM_RATIO	260,951	1.137	0.794	1.305	0.414	1.403
LEVERAGE	261,031	0.226	0.192	0.206	0.035	0.359
EARN_SURPRISE	260,971	0.041	0.005	0.333	-0.036	0.055
NI_LOSS	266,295	0.300	0.000	0.458	0.000	1.000
AGE	261,245	12.201	10.637	8.041	5.919	16.266

Panel C: Comparison of our sample with the Datastream population by Language Group

Country	Our Sample				Datastream Population			
	N	LN_MVE	ROA	%ADR	N	LN_MVE	ROA	%ADR
Non-English	34,136	12.643	0.028	0.016	164,624	11.538	0.022	0.023
English-Speaking	53,472	11.052	0.021	0.003	101,671	10.754	0.017	0.015

Panel D: Median Text Characteristics by Year

Year	N	NUM_WORDS	BOILERPLATE	XUS_COMP	US_COMP	FOG	LN_MVE
1998	1476	11611.5	5.86	0.62	0.62	18.3	12.99
1999	2718	11778.5	6.51	0.64	0.61	18.5	12.66
2000	3821	12038	6.85	0.64	0.61	18.8	12.26
2001	4953	12017	7.21	0.65	0.61	19.1	11.58
2002	5581	13340	7.52	0.66	0.62	19.1	11.34
2003	6160	14399.5	7.87	0.66	0.62	19.3	11.32
2004	6482	15603	7.91	0.67	0.62	19.3	11.85
2005	6794	17873	7.48	0.69	0.64	19.5	12.01
2006	8235	18072	7.59	0.7	0.64	19.6	11.66
2007	8897	19874	7.1	0.71	0.64	19.9	11.80
2008	8830	21060	6.89	0.72	0.65	20	11.36
2009	9460	20984.5	6.72	0.72	0.65	19.9	11.01
2010	9003	22298	6.74	0.72	0.65	20	11.33
2011	5198	24860.5	5.42	0.73	0.65	20.4	11.85

Panel E: Median Text Characteristics by Country

Country	N	NUM_WORDS	BOILERPLATE	XUS_COMP	US_COMP	FOG	LN_MVE
ARGENTINA	147	4475.09	1.99	-0.03	0.02	-0.54	12.99
AUSTRALIA	9398	2184.36	1.51	0.01	-0.01	1.46	10.48
AUSTRIA	576	2221.18	-2.15	0	0	-1.46	12.60
BELGIUM	743	-33.8	-2.82	0.01	0	-1.12	12.82
BRAZIL	504	-9716.75	-2.82	0	0.03	0.57	13.49
CANADA	11653	-4741.51	-5.1	0	0.05	0.05	10.75
CHILE	121	2327.67	-2.47	0	0.02	-1.23	14.06
CHINA	1446	6378.58	5.23	0.01	0	1.92	11.48
CZECH REPUBLIC	169	6425.21	-3.4	-0.02	0	-1.46	11.62
DENMARK	969	-2161.46	-2.92	0.02	0	-0.8	12.02
FINLAND	1161	344.69	-2.46	0.02	0.01	-1.5	12.13
FRANCE	1934	7305.23	-2.75	0	-0.01	-0.02	13.29
GERMANY	3692	5686.8	-0.55	-0.01	0	-1.38	12.00
GREECE	109	-6648.29	-6.29	0.01	-0.03	-1.97	12.92
HONG KONG	2147	1751.43	3.82	0.03	-0.02	1.95	11.96
INDIA	6289	-912.42	1.21	-0.01	0	-2.24	11.42

48

INDONESIA	1112	4524.51	2.52	-0.17	-0.14	1.21	10.97
IRELAND	519	-1681.75	-0.6	0.05	-0.01	0.13	11.69
ISRAEL	467	2637.64	0.06	0.01	0.05	0.16	12.29
ITALY	1283	10325.32	1.67	-0.01	-0.01	-1.44	13.27
JAPAN	5700	-13302.99	1.83	-0.01	0.02	0.46	14.09
KOREA (SOUTH)	443	-11528.83	-0.78	-0.01	0.02	0.42	14.07
LUXEMBOURG	190	-3844.57	-5.37	0.03	-0.01	-0.94	13.23
MEXICO	354	-9618.87	-1.51	0.02	0.06	-0.87	13.85
NETHERLANDS	1074	1486.33	-4.14	0	0	-1.65	12.79
NEW ZEALAND	983	-3669.1	-1.74	0.02	-0.02	0.28	11.10
NORWAY	1124	821.62	-3.88	0.02	0.01	-1.6	11.97
PAKISTAN	590	-4075.44	3.9	0.04	0.03	-0.85	11.57
PHILIPPINES	1361	5866.55	1.8	0	0.05	-0.63	10.36
POLAND	511	2493.85	-1.65	-0.01	-0.04	-0.84	12.62
PORTUGAL	200	16101.53	2.92	0.01	0	0.15	14.25
RUSSIA	212	-4344.28	-3.82	0.01	-0.01	1.41	14.62
SINGAPORE	4438	3506.21	2.61	0.04	0	1.35	10.80
SOUTH AFRICA	2048	3253.6	0.98	0.04	-0.01	2	11.83
SPAIN	596	6827.25	-3.11	0	-0.03	-0.54	14.50
SRI LANKA	636	-599.33	0.6	0.04	0	-2.07	9.37
SWEDEN	1875	2830.04	-2.13	-0.01	-0.01	-1.11	12.21
SWITZERLAND	1909	-3227.41	-3.18	0.02	0	-1.14	12.84
TAIWAN	528	-11347.27	-0.35	-0.02	-0.01	0.08	14.02
THAILAND	3701	-10589.63	-0.08	-0.02	-0.02	0.09	10.45
TURKEY	650	-1971.68	-1.39	-0.01	-0.01	-0.48	12.69
UNITED KINGDOM	14046	-1287.58	1.18	0.03	0	0.23	11.33

Panel A reports summary statistics for the variables used in our paper. Panel B reports summary statistics for the entire Datastream population in the same country-years. Panel C compares median size, profitability and percent of firms with an ADR for our sample and the entire Datastream population, partitioning by whether the firm's country of origin is English-speaking or not. Panel D reports median textual characteristics for each year in our sample. Panel E reports median textual characteristics by country. In order to adjust for sample composition, and for ease of comparison across groups, the text variables in Panel E have been adjusted for size and year. Inferences about the trends and patterns in Panels D and E are robust to controlling (or not controlling) for a wide range of fundamentals such as size, industry, year, performance, leverage, Big 5 auditor, etc.

Table 3. Correlation Matrices

Panel A. Text Attributes and Select Determinants

		1	2	3	4	5	6	7	8	9	10	11
LN_NBWORDS	1		-0.106	0.290	0.260	-0.059	0.463	0.136	0.231	0.370	0.029	0.067
BOILERPLATE	2	-0.105		0.116	-0.136	0.203	-0.146	0.035	-0.134	-0.081	-0.064	-0.023
XUS_COMP	3	0.301	0.109		0.590	0.232	-0.098	-0.065	0.040	0.324	-0.018	0.004
US_COMP	4	0.271	-0.157	0.521		0.027	0.112	-0.012	0.062	0.136	0.083	0.052
FOG	5	-0.061	0.212	0.259	0.020		-0.229	-0.172	-0.035	0.102	-0.037	-0.014
LN_MVE	6	0.456	-0.153	-0.111	0.118	-0.244		0.306	0.239	0.044	0.066	0.134
ROA	7	0.133	0.049	-0.047	-0.031	-0.155	0.318		0.101	-0.048	-0.009	0.021
BIG5	8	0.234	-0.138	0.040	0.057	-0.041	0.247	0.115		0.063	0.019	0.021
IFRS	9	0.372	-0.081	0.355	0.148	0.099	0.043	-0.026	0.063		-0.078	-0.027
US_GAAP	10	0.029	-0.065	-0.025	0.086	-0.041	0.054	-0.015	0.019	-0.078		0.086
ADR	11	0.065	-0.025	0.002	0.055	-0.015	0.113	0.029	0.021	-0.027	0.086	

Panel B. Text Attributes and Outcome Variables

		1	2	3	4	5	6	7	8	9
LN_NBWORDS	1		-0.106	0.290	0.260	-0.059	-0.333	-0.281	0.279	0.218
BOILERPLATE	2	-0.105		0.116	-0.136	0.203	0.180	0.072	-0.128	-0.111
XUS_COMP	3	0.301	0.109		0.590	0.232	0.122	0.135	-0.030	0.080
US_COMP	4	0.271	-0.157	0.521		0.027	-0.115	-0.024	0.067	0.070
FOG	5	-0.061	0.212	0.259	0.020		0.295	0.257	-0.108	-0.107
ZERO_RETURN	6	-0.364	0.147	0.110	-0.140	0.313		0.675	-0.384	-0.149
LN_BIDASK	7	-0.374	0.148	0.176	-0.085	0.312	0.782		-0.274	-0.146
ANALYST	8	0.279	-0.161	-0.065	0.078	-0.133	-0.457	-0.558		0.229
INST_OWN	9	0.229	-0.151	0.047	0.078	-0.143	-0.232	-0.266	0.322	

Pearson/Spearman Correlations given above/below the diagonal.

50

Table 4. Economic Determinants of Textual Attributes

Panel A. Mixed Model

VARIABLES	(1) LN_NBWORDS	(2) BOILERPLATE	(3) XUS_COMP	(4) US_COMP	(5) FOG	(6) FACTOR1	(7) FACTOR2
IFRS	0.220***	-0.682***	0.0197***	0.0141***	0.193***	0.376***	-0.00387
	(36.07)	(-16.27)	(26.25)	(20.92)	(10.70)	(40.66)	(-0.554)
US_GAAP	0.0881***	-1.585***	0.0164***	0.0370***	0.00892	0.489***	-0.174***
	(3.903)	(-11.23)	(5.330)	(11.56)	(0.149)	(10.79)	(-6.678)
ADR	0.117***	-0.0296	0.00969**	0.0149***	0.192*	0.316***	0.0269
	(2.959)	(-0.120)	(2.214)	(3.491)	(1.903)	(4.957)	(0.745)
BIG5	0.0676***	-0.0889***	0.00144**	-0.000526	0.0215	0.0469***	-0.00673
	(13.64)	(-2.875)	(2.493)	(-1.047)	(1.551)	(6.250)	(-1.283)
LN_MVE	0.115***	-0.0652***	-0.00508***	-0.00149***	-0.0630***	0.0360***	-0.0634***
	(67.13)	(-5.537)	(-21.86)	(-6.861)	(-12.33)	(12.55)	(-30.26)
BM_RATIO	0.0517***	0.0147	-0.00209***	0.000153	-0.0141**	0.0229***	-0.0240***
	(26.80)	(1.203)	(-9.012)	(0.720)	(-2.540)	(7.667)	(-10.74)
LEVERAGE	0.238***	0.256***	-0.0181***	-0.00894***	-0.566***	-0.0287	-0.195***
	(18.15)	(2.969)	(-10.93)	(-6.002)	(-14.26)	(-1.381)	(-12.53)
ROA	0.0306***	0.0566	-0.0203***	-0.0169***	-0.255***	-0.311***	-0.276***
	(3.891)	(1.076)	(-6.640)	(-5.111)	(-10.97)	(-6.222)	(-7.336)
AGE	-0.00150***	0.0111***	0.000273***	0.000617***	-0.0159***	0.00574***	-0.00205***
	(-3.515)	(4.004)	(6.044)	(13.63)	(-12.75)	(9.571)	(-4.282)
NI_LOSS	-0.0159***	-0.0960***	0.000414	0.00211***	0.0885***	0.0113*	0.00710
	(-3.947)	(-3.790)	(0.825)	(4.878)	(8.114)	(1.759)	(1.597)
EARN_SURPRISE	0.00899**	-0.0315	0.00146***	0.00268***	0.106***	0.0382***	0.0108**
	(2.097)	(-1.189)	(2.652)	(5.611)	(8.597)	(5.449)	(2.279)
COMP Match Quality Measures	N	N	Y	Y	N	Y	Y
NUM_WORDS_1-5 Included	N	Y	Y	Y	N	N	N
Year Fixed Effects	Y	Y	Y	Y	Y	Y	Y
Firm Random Effects	Y	Y	Y	Y	Y	Y	Y
Observations	83,241	81,626	78,955	80,003	81,556	74,350	74,350
R-Squared	0.398	0.031	0.248	0.274	0.110	0.198	0.159

Panel B. Firm and Year Fixed Effects

VARIABLES	(1) LN_NBWORDS	(2) BOILERPLATE	(3) XUS_COMP	(4) US_COMP	(5) FOG	(6) FACTOR1	(7) FACTOR2
IFRS	0.170***	-0.783***	0.0181***	0.0172***	0.160***	0.378***	-0.0299***
	(21.68)	(-14.73)	(18.55)	(20.08)	(6.944)	(30.78)	(-3.319)
US_GAAP	0.0650**	-1.453***	0.0175***	0.0336***	0.0728	0.459***	-0.132***
	(2.184)	(-7.532)	(4.060)	(7.752)	(0.932)	(7.057)	(-3.893)
ADR	0.000288	0.446	-0.00498	-0.00618	0.122	0.00540	0.114*
	(0.00320)	(0.924)	(-0.461)	(-0.682)	(0.617)	(0.0330)	(1.797)
BIG5	0.0214***	0.0381	-0.000395	-0.00126**	0.00645	0.000553	0.00269
	(3.356)	(0.991)	(-0.524)	(-1.974)	(0.370)	(0.0559)	(0.411)
LN_MVE	0.0533***	-0.00900	-0.00193***	-0.000383	-0.00664	0.0251***	-0.0277***
	(16.92)	(-0.469)	(-4.918)	(-1.080)	(-0.744)	(4.700)	(-7.736)
BM_RATIO	0.0249***	0.0147	-0.000388	0.000784***	0.0106	0.0218***	-0.00927***
	(9.768)	(0.935)	(-1.229)	(2.780)	(1.437)	(5.262)	(-3.160)
LEVERAGE	0.146***	0.142	-0.00461*	0.00140	-0.201***	0.0881***	-0.0879***
	(7.959)	(1.230)	(-1.924)	(0.682)	(-3.755)	(2.927)	(-4.110)
ROA	0.00362	-0.0743	-0.00559	-0.00662	-0.125***	-0.115	-0.253***
	(0.374)	(-1.201)	(-1.335)	(-1.321)	(-4.438)	(-1.387)	(-4.634)
AGE	0.0614	-0.547**	0.00111	-0.000740	0.152	0.0620	-0.0131
	(1.549)	(-2.044)	(0.213)	(-0.158)	(1.194)	(0.895)	(-0.295)
NI_LOSS	0.00343	0.0113	-0.000443	0.000978*	0.0372***	0.00758	0.00245
	(0.731)	(0.386)	(-0.742)	(1.937)	(2.923)	(0.991)	(0.470)
EARN_SURPRISE	0.00334	0.0345	0.00111*	0.00167***	0.0708***	0.0231***	0.0115**
	(0.680)	(1.137)	(1.705)	(3.015)	(5.007)	(2.775)	(2.089)
COMP Match Quality Measures	N	N	Y	Y	N	Y	Y
NUM_WORDS_1-5 Included	N	Y	Y	Y	N	N	N
Firm & Year Fixed Effects	Y	Y	Y	Y	Y	Y	Y
Observations	83,241	81,626	78,955	80,003	81,556	74,350	74,350
Adjusted R-Squared	0.807	0.774	0.730	0.732	0.788	0.733	0.812

The tests reported in this table examine the economic determinants of textual attributes. Panel A presents results from a mixed model with year fixed effects and firm random effects; Panel B includes firm and year fixed effects. All continuous, non-logarithmic variables are truncated at the 1st and 99th percentiles, except for the factors, which were formed from truncated variables. We used lagged values of LN_MVE, BM_RATIO, and LEVERAGE but current values of the other control variables which should be reflected in the current annual report. . We include the first five powers of NUM_WORDS to remove mechanical effects in boilerplate and comparability due to report length (Brown and Tucker 2011). We also control for the peer match quality of the COMP measures (see the variable definitions section).

Robust t-statistics, clustered by firm, are reported in parentheses.

*** $p<0.01$, ** $p<0.05$, * $p<0.1$

Table 5. Factor Analysis of Textual Characteristics

	Factor Pattern		Factor Pattern: Varimax Rotation	
	Factor 1	Factor 2	Factor 1	Factor 2
LN_NBWORDS	0.398	-0.227	0.429	-0.161
BOILERPLATE	-0.038	0.470	-0.112	0.459
XUS_COMP	0.755	0.230	0.709	0.347
US_COMP	0.741	-0.177	0.760	-0.057
FOG	0.143	0.460	0.068	0.476
Eigenvalue	1.300	0.569	-	-

This table provides the results of a maximum likelihood factor analysis of our five main textual measures. We present both the raw factor patterns as well as the patterns generated after a varimax rotation of the factors. We retain two factors from the factor analysis because these two are able to explain at least 90% of the observed variance of the original variables. These two factors would also be retained in a Cattell scree test.

Table 6. Economic Outcomes

VARIABLES	(1) ZERO_RETURN	(2) LN_BIDASK	(3) ANALYST	(4) INST_OWN
FACTOR1	-1.190***	-0.000741	0.146***	0.421***
	(-6.828)	(-1.554)	(2.699)	(3.336)
FACTOR2	2.226***	0.00288***	-0.463***	-0.793***
	(7.420)	(3.315)	(-5.183)	(-3.742)
IFRS	-2.667***	0.00179*	0.907***	-0.312
	(-7.222)	(1.787)	(7.800)	(-1.150)
US_GAAP	-3.402***	-0.00459***	0.717	0.967
	(-3.777)	(-2.960)	(1.632)	(0.826)
ADR	-7.559	0.00247	-3.263	-1.387
	(-1.227)	(0.565)	(-1.605)	(-0.668)
BIG5	-0.808***	-0.00170**	0.0172	0.334
	(-2.691)	(-2.067)	(0.210)	(1.626)
LN_MVE	-4.079***	-0.00739***	1.124***	1.758***
	(-21.14)	(-11.00)	(19.24)	(10.93)
BM_RATIO	-1.213***	-0.00327***	0.298***	0.688***
	(-7.505)	(-5.840)	(6.239)	(5.929)
LEVERAGE	0.681	0.00426	-0.374	-1.474*
	(0.622)	(1.073)	(-1.222)	(-1.864)
NI_LOSS	1.620***	0.00444***	-0.289***	-0.483***
	(6.483)	(5.160)	(-4.230)	(-2.840)
ROA	-8.111***	-0.0466***	-1.250	1.771
	(-2.788)	(-4.602)	(-1.632)	(0.769)
COMP Match Quality Measures	Y	Y	Y	Y
Firm & Year Fixed Effects	Y	Y	Y	Y
Observations	53,912	42,842	35,659	35,294
Adjusted R-Squared	0.852	0.800	0.791	0.746

This table examines the effects of our text factors on several economic outcomes. All continuous, non-logarithmic variables are truncated at the 1st and 99th percentiles except for the factors, which were formed from truncated variables. We lag our independent variables to mitigate endogeneity problems. We also control for the quality of the peer matches of the COMP measures.

Robust t-statistics, clustered by firm, are reported in parentheses

*** $p<0.01$, ** $p<0.05$, * $p<0.1$

Table 7. Difference-in-Differences Test of the Effect of IFRS on Textual Attributes

VARIABLES	(1) LN_NBWORDS	(2) BOILERPLATE	(3) XUS_COMP	(4) IFRS_COMP	(5) US_COMP	(6) FOG	(7) FACTOR1	(8) FACTOR2
POST x MANDATORY	0.201***	-0.578***	0.0180***	0.0466***	0.0202***	0.00767	0.433***	-0.0466***
	(18.46)	(-7.122)	(12.59)	(26.40)	(16.17)	(0.241)	(24.98)	(-3.815)
US_GAAP	0.0808**	-1.503***	0.0159***	0.0135**	0.0318***	0.0207	0.427***	-0.146***
	(2.017)	(-5.554)	(2.944)	(2.559)	(5.647)	(0.210)	(4.954)	(-3.199)
ADR	0.0711	-0.305	-0.00685	0.00664	-0.00214	0.0633	0.0512	0.0411
	(0.576)	(-0.760)	(-0.506)	(0.516)	(-0.211)	(0.302)	(0.248)	(1.026)
BIG5	0.00914	0.00285	-0.00123	0.00150	-0.00268***	0.00956	-0.0275**	0.00384
	(1.066)	(0.0521)	(-1.279)	(1.295)	(-3.229)	(0.394)	(-2.157)	(0.425)
LN_MVE	0.0507***	-0.0430	-0.00272***	-0.000791	-0.000246	-0.0419***	0.0190***	-0.0356***
	(11.26)	(-1.493)	(-5.135)	(-1.178)	(-0.511)	(-3.207)	(2.609)	(-7.156)
BM_RATIO	0.0236***	0.0150	-0.00107**	-0.000365	0.00106***	0.00153	0.0194***	-0.0113***
	(6.372)	(0.603)	(-2.397)	(-0.663)	(2.640)	(0.137)	(3.239)	(-2.645)
LEVERAGE	0.134***	0.0812	-0.00575*	-0.000316	0.00500*	-0.333***	0.0943**	-0.118***
	(5.138)	(0.478)	(-1.750)	(-0.0772)	(1.784)	(-4.296)	(2.278)	(-3.997)
ROA	-0.00496	0.0594	-0.00466	-0.0110	-0.0112	-0.0721	-0.178	-0.197**
	(-0.325)	(0.558)	(-0.694)	(-1.425)	(-1.326)	(-1.536)	(-1.321)	(-2.511)
AGE	0.0631	-0.243	0.00475	0.00832	0.000644	0.388**	0.0688	0.0754
	(1.053)	(-0.633)	(0.636)	(0.836)	(0.0925)	(1.991)	(0.656)	(1.254)
NI_LOSS	0.00885	-0.0523	-0.000560	-0.00146	0.00103	0.0588***	0.0119	-0.00222
	(1.330)	(-1.232)	(-0.680)	(-1.409)	(1.465)	(3.246)	(1.115)	(-0.304)
EARN_SURPRISE	0.0159**	-0.0322	0.000452	0.00136	0.00124	0.0546**	0.0250**	-0.00334
	(2.249)	(-0.662)	(0.475)	(1.172)	(1.440)	(2.517)	(2.002)	(-0.394)
COMP Match Quality Measures	N	N	Y	Y	Y	N	Y	Y
NUM_WORDS_1-5 Included	N	Y	Y	Y	Y	N	N	N
Firm & Year Fixed Effects	Y	Y	Y	Y	Y	Y	Y	Y
Observations	39,027	38,354	37,185	35,678	37,766	38,219	35,159	35,159
Adjusted R-Squared	0.815	0.731	0.733	0.697	0.727	0.782	0.737	0.806

The tests in this table use a difference-in-differences design to examine the effect of mandatory IFRS adoption on textual disclosure. MANDATORY is an indicator variable equal to one if a firm adopted IFRS in 2005 and was in a country that mandated IFRS for the first time in that year. POST is an indicator variable equal to one if the fiscal year is post 2005. POST x MANDATORY can be interpreted as the effect of mandatory IFRS adoption on disclosure. POST and MANDATORY are excluded because they are perfectly collinear with our firm and year fixed effects. IFRS_COMP is the textual comparability with non-US peers *within* IFRS adoption groups so that (non-)adopters are only compared with other (non-)adopters. Our sample includes only those firms which have available data in 2004 and 2005, the year before and of IFRS adoption.

All continuous, non-logarithmic variables are truncated at the 1st and 99th percentiles, except for the factors, which were formed from truncated variables. As in Table 4, LN_MVE, BM_RATIO, and LEVERAGE are lagged. We include the first five powers of NUM_WORDS to remove mechanical effects in boilerplate and comparability due to report length. We also control for the peer match quality of the COMP measures.

Robust t-statistics, clustered by firm, are reported in parentheses.

*** p<0.01, ** p<0.05, * p<0.1

Table 8. Significant Changes in Disclosure Topics around IFRS Adoption

	Topic Keywords	Post-IFRS Change	Topic Weight
1	*options, shares, share, option, price, granted, exercise, date, plan, period, grant, number, issued, company, employees, total, employee, average, exercised*	0.006 *** (12.96)	0.218
2	*interest, fair, rate, instruments, derivative, financial, hedge, contracts, net, cash, income, assets, hedging, risk, derivatives, debt, hedges, losses, liabilities,*	0.009 *** (16.18)	0.156
3	*group, recognised, assets, tax, income, liabilities, cash, amount, impairment, financial, profit, cost, net, costs, fair, equity, plant, equipment, property*	0.033 *** (21.43)	0.110
4	*financial, year, group, directors, statements, company, cash, report, statement, share, tax, assets, limited, capital, march, recognised, ended, income, balance*	0.039 *** (20.08)	0.043
5	*group, year, performance, committee, executive, share, scheme, profit, board, remuneration, business, pension, directors, shares, tax, years, schemes, total, annual,*	0.009 *** (11.94)	0.043
6	*directors, financial, remuneration, year, report, board, company, director, management, limited, tax, recognised, risk, accounting, performance, key, executive, reporting, cash,*	0.026 *** (17.25)	0.037
7	*board, supervisory, financial, management, year, report, company, members, statements, thousand, cash, capital, annual, tax, meeting, liabilities, corporate, risks, risk*	0.008 *** (12.05)	0.028
8	*group, limited, year, company, directors, share, investment, profit, continued, subsidiaries, interest, interests, assets, note, loss, china, shares, recognised, capital*	0.016 *** (12.34)	0.023
9	*company, financial, year, group, shares, directors, director, companies, share, tax, recognised, statements, limited, interest, profit, subsidiary, subsidiaries, assets, loss*	0.014 *** (13.60)	0.017

This table lists the top 20 keywords for LDA topics that were found to dramatically increase in annual reports after mandatory IFRS adoption. It also provides the values and t-statistics for the coefficient β1 from the following set of regressions:

$$Topic\_N_{it} = \alpha + \beta_1 Post \times Mandatory_{it} + \beta_2 Ln\_Words_{it-1} + \gamma_t + \delta_i + \varepsilon_{it}$$

where Topic_N is the topic loading for the Nth topic. β1 is the effect of Mandatory IFRS adoption on the proportion of the annual report that is composed of Topic_N. In order to focus on topics that show the most statistically and economically significant changes around IFRS adoption, we report only those topics where β1 has a robust, firm-clustered t-statistic of 10 or greater. The current specification includes firm and year fixed effects and a control for lagged length to ensure that static firm- or year-specific factors and changes in document length do not drive the observed changes in topic proportions. These topics also exhibit significant increases when examined in specifications omitting length, including current length, or including the full set of control variables from Table 4. Topic Weight is the Dirichlet parameter for the given topic, which is roughly proportional to the overall portion of the collection of documents assigned to a given topic (the maximum topic weight in our set of 100 topics is 0.318 and the minimum is 0.005).

Table 9. Instrumental Variables Test of Economic Outcomes

Panel A. Instrument Stength

VARIABLES	(1) FACTOR1	(2) FACTOR2
IFRS	0.205***	-0.0433***
	(13.06)	(-3.940)
MEAN_PEER_FACTOR1	0.823***	0.0161*
	(57.58)	(1.688)
MEAN_PEER_FACTOR2	0.0177	0.932***
	(1.161)	(91.08)
MEAN_PEER_ADR	-0.00932	0.0670
	(-0.0517)	(0.580)
MEAN_PEER_BIG5	-0.130***	0.0353*
	(-4.412)	(1.848)
CONSTANT	-0.0695***	-0.0624***
	(-4.251)	(-5.865)
Observations	26,562	26,562
Prob > F-Statistic	0	0
Adjusted R-Squared	0.378	0.550

Panel B. Second Stage Regressions, Using Mandatory IFRS Adoption as an Instrument

VARIABLES	(1) ZERO_RET	(2) LN_BIDASK	(3) ANALYST	(4) INST_OWN
IV_FACTOR1	-9.755***	-0.00246	2.315***	0.432
	(-8.698)	(-1.021)	(5.637)	(0.545)
IV_FACTOR2	22.07***	0.0289***	-5.793***	-11.71***
	(6.292)	(3.621)	(-5.170)	(-3.522)
US_GAAP	4.577***	0.000502	-1.662***	-0.675
	(2.950)	(0.195)	(-2.905)	(-0.450)
ADR	9.776*	0.00582	-7.142***	-1.811
	(1.709)	(0.612)	(-2.750)	(-0.693)
BIG5	-0.988**	-0.00202**	-0.00808	0.340
	(-2.448)	(-2.068)	(-0.0500)	(1.161)
LN_MVE	-3.255***	-0.00612***	0.981***	1.512***
	(-10.94)	(-7.846)	(12.63)	(5.700)
BM_RATIO	-0.809***	-0.00247***	0.191***	0.763***
	(-3.575)	(-3.421)	(2.944)	(4.174)
LEVERAGE	3.133*	0.00377	-0.449	-2.419**
	(1.910)	(0.679)	(-0.975)	(-2.341)
NI_LOSS	1.874***	0.00413***	-0.257**	-0.662**
	(5.604)	(3.364)	(-2.437)	(-2.335)
ROA	0.671	-0.0369***	-3.398***	-0.519
	(0.147)	(-3.530)	(-3.273)	(-0.148)
COMP Match Quality Measures	Y	Y	Y	Y
Firm & Year Fixed Effects	Y	Y	Y	Y
Observations	24,966	20,020	17,019	18,495
R-Squared	0.269	0.191	0.208	0.050

Panel C. Second Stage Regressions, Using Only Peer Variables as Instruments

VARIABLES	(1) ZERO_RET	(2) LN_BIDASK	(3) ANALYST	(4) INST_OWN
IV_FACTOR1	-11.35***	-0.00363	1.985***	3.377***
	(-7.974)	(-1.156)	(3.478)	(2.753)
IV_FACTOR2	17.58***	0.0191***	-4.753***	-9.775***
	(7.249)	(2.988)	(-6.385)	(-5.142)
IFRS	2.051***	0.00370**	-0.151	-1.911***
	(2.898)	(2.276)	(-0.517)	(-3.757)
US_GAAP	2.051***	0.00370**	-0.151	-1.911***
	(2.898)	(2.276)	(-0.517)	(-3.757)
ADR	4.340***	0.000986	-1.117*	-2.586*
	(2.911)	(0.359)	(-1.855)	(-1.923)
BIG5	11.20	0.0104	-7.053***	-2.413
	(1.589)	(1.634)	(-3.274)	(-0.831)
LN_MVE	-0.782***	-0.00128*	-0.0227	0.359
	(-3.053)	(-1.763)	(-0.313)	(1.620)
BM_RATIO	-3.424***	-0.00671***	0.982***	1.461***
	(-19.89)	(-13.14)	(15.80)	(9.177)
LEVERAGE	-0.900***	-0.00317***	0.183***	0.523***
	(-6.069)	(-5.930)	(3.678)	(4.178)
NI_LOSS	3.070***	0.00503	-0.599	-2.315***
	(3.366)	(1.257)	(-1.611)	(-2.651)
ROA	1.814***	0.00454***	-0.279***	-0.523***
	(7.442)	(5.406)	(-4.275)	(-3.060)
COMP Match Quality Measures	Y	Y	Y	Y
Country, Industry, Year Fixed Effects	Y	Y	Y	Y
Observations	47,334	37,689	31,291	31,401
R-Squared	0.291	0.271	0.219	0.050

This table uses instrumental variables (IV) analysis to examine the effects of our text factors on several economic outcomes. Panel A provides evidence on the strength of our instruments by showing regressions of Factor1 and Factor2 on our exogenous variables (the actual first stage regressions also include the other control variables and vary in sample composition across specifications according to data restrictions). Panel B provides second stage estimates of the effect of exogenous changes in disclosure on outcomes using mandatory IFRS adoption as the main instrument. Panel C provides second stage estimates of the effect of disclosure on outcomes for the full sample of firms, using only the peer variables as instruments. All continuous, non-logarithmic variables are truncated at the 1st and 99th percentiles except for the factors, which were formed from truncated variables. We lag our independent variables to mitigate endogeneity problems. We also control for the quality of the peer matches of the COMP measures. Robust t-statistics clustered by firm reported in parentheses in Panel A. Bootstrapped t-statistics reported in Panels B and C.

*** p<0.01, ** p<0.05, * p<0.1

Table 10. The Effect of Textual Attributes on the Benefits of IFRS Adoption

Panel A. OLS Analysis

VARIABLES	(1) ZERO_RETURN	(2) LN_BIDASK	(3) ANALYST	(4) INST_OWN
POST x MANDATORY x HBen	-3.413***	-0.00200	0.870***	1.447**
	(-3.655)	(-1.419)	(3.685)	(2.267)
POST x MANDATORY x LBen	2.048**	-0.00148	-1.101***	-1.755**
	(2.126)	(-0.758)	(-4.339)	(-2.264)
POST x MANDATORY	-5.516***	-0.00387***	1.163***	0.0603
	(-8.500)	(-2.603)	(6.000)	(0.145)
US_GAAP	-3.567***	-0.00730***	0.476	1.156
	(-4.755)	(-4.853)	(0.997)	(0.737)
ADR	-6.311	0.00157	-2.833	-4.379***
	(-0.928)	(0.328)	(-1.294)	(-5.334)
BIG5	-0.681*	-0.00204*	0.0147	0.159
	(-1.674)	(-1.876)	(0.133)	(0.567)
LN_MVE	-4.074***	-0.00683***	1.291***	1.977***
	(-15.24)	(-7.695)	(16.50)	(8.834)
BM_RATIO	-0.992***	-0.00264***	0.371***	0.955***
	(-4.164)	(-3.158)	(5.218)	(5.522)
LEVERAGE	0.218	0.00207	-0.0893	-1.674
	(0.137)	(0.374)	(-0.206)	(-1.524)
NI_LOSS	1.732***	0.00382***	-0.301***	-0.667***
	(4.824)	(3.337)	(-3.111)	(-2.902)
ROA	-6.013	-0.0346**	-2.323**	3.014
	(-1.436)	(-2.382)	(-1.980)	(0.877)
COMP Match Quality Measures	Y	Y	Y	Y
Firm & Year Fixed Effects	Y	Y	Y	Y
Observations	24,198	19,361	16,766	18,149
Adjusted R-Squared	0.827	0.788	0.793	0.745

Panel B. Instrumental Variables Analysis

VARIABLES	(1) ZERO_RETURN	(2) LN_BIDASK	(3) ANALYST	(4) INST_OWN
IV_POST x MANDATORY x HBen	-13.95***	-0.00127	2.044***	-1.175
	(-3.617)	(-0.199)	(2.911)	(-0.543)
IV_POST x MANDATORY x LBen	7.510**	-0.00850	-1.594*	1.616
	(2.029)	(-1.203)	(-1.700)	(0.567)
POST x MANDATORY	-2.795*	-0.00277	0.791**	0.319
	(-1.917)	(-1.095)	(2.504)	(0.380)
US_GAAP	-1.447	-0.00813***	0.215	1.950
	(-1.419)	(-4.652)	(0.461)	(1.200)
ADR	-6.762	0.00219	-2.672	-4.764***
	(-1.077)	(0.510)	(-0.941)	(-4.429)
BIG5	-0.789**	-0.00194*	0.0265	0.103
	(-2.354)	(-1.871)	(0.268)	(0.395)
LN_MVE	-4.103***	-0.00682***	1.290***	1.954***
	(-17.17)	(-7.943)	(20.85)	(8.996)
BM_RATIO	-0.990***	-0.00264***	0.370***	0.943***
	(-4.209)	(-3.438)	(5.360)	(6.852)
LEVERAGE	0.345	0.00191	-0.116	-1.653
	(0.210)	(0.416)	(-0.233)	(-1.608)
NI_LOSS	1.797***	0.00376***	-0.304***	-0.629***
	(5.570)	(4.183)	(-3.591)	(-3.182)
ROA	-7.163*	-0.0345***	-2.045*	2.514
	(-1.917)	(-2.821)	(-1.816)	(0.767)
COMP Match Quality Measures	Y	Y	Y	Y
Firm & Year Fixed Effects	Y	Y	Y	Y
Observations	24,198	19,361	16,766	18,149
R-Squared	0.288	0.209	0.375	0.056

The tests in this table examine heterogeneity in the benefits of mandatory IFRS adoption, in particular whether firms with large increases in beneficial disclosure (Factor1) received greater benefits of adoption, and whether firms with large increases in unbeneficial disclosure (Factor2) received smaller benefits of adoption. The base specification is a difference-in-difference design where POST x MANDATORY (defined as in Table 7) is the benefit for firms who mandatorily adopted IFRS. HBen (LBen) is an indicator variable equal to one if the change in FACTOR1 (FACTOR2) in the year of IFRS adoption was in the top quintile for all firms. Thus the interaction of POST x MANDATORY with HBen (LBen) is the incremental benefit of mandatory adoption for firms that showed high (low) quality changes in their disclosure around adoption. Panel A reports OLS results. Panel B reports the results of an instrumental variables regression where 3-digit ICB industry dummies and their interactions with POST x MANDATORY are used as instruments. POST, MANDATORY, HBen and LBen are excluded because they are perfectly collinear with our firm and year fixed effects.

All continuous, non-logarithmic variables are truncated at the 1st and 99th percentiles. All control variables are lagged to mitigate endogeneity problems. We also control for the quality of the peer matches of the COMP measures.

Robust t-statistics clustered by firm reported in Panel A. Bootstrapped t-statistics reported in Panel B.

*** p<0.01, ** p<0.05, * p<0.1