Volume 23, Number 1 – March 2021
China Accounting and Finance Review
中 国 会 计 与 财 务 研 究
2021 年 3 月 第 23 卷 第 1 期

Integrated Reporting and the Informativeness of Annual Reports: Does Textual Coherence Matter?*

Nishant Agarwal[1]

Received 26th of June 2020 Accepted 2nd of February 2021

Abstract

This study examines the influence of textual coherence on the informativeness of annual integrated reports in South Africa. Using latent semantic analysis (LSA) to measure coherence, I demonstrate that the informativeness of integrated reports is positively associated with their coherence. Further analysis reveals that coherence mitigates the negative effects of linguistic complexity, captured by the Fog Index, on report informativeness and stock price delay. My findings contribute to the growing literature on integrated reporting (IR) by documenting a potential benefit for investors and also suggest that the Fog Index or other traditional measures of linguistic complexity may not be sufficient to gauge the consequences of IR adoption on the informativeness of annual reports.

***Keywords*:** Disclosures, Capital Markets, Complexity, Integrated Reporting, Latent Semantic Analysis, Coherence

***JEL classification*:** M41

* **Acknowledgements:** I sincerely thank members of my dissertation committee, Sanjay Kallapur (Indian School of Business), Ranjani Krishnan (Michigan State University), and Hariom Manchiraju (Indian School of Business), for their valuable inputs across various stages of this paper. I would also like to extend my gratitude to the discussants and participants at seminars at the Indian School of Business, the EAA Talent Workshop (2017), AAA Rookie Camp (2017), KU Leuven (2018), Copenhagen Business School (2018), and UWA Business School (2018). Last but not least, I would like to thank my colleagues at the Indian School of Business for their comments and critiques at various PhD seminars. I take complete responsibility for any errors.

[1] Department of Accounting and Finance, UWA Business School, University of Western Australia, Perth, Australia. Email: nishant.agarwal@uwa.edu.au.

I. Introduction

A growing body of literature documents the inability of annual reports to clearly communicate relevant information to investors (Dyer *et al.*, 2017; Guay *et al.*, 2016; KPMG, 2011; Li, 2008; SEC, 2013). Textual attributes, such as the Fog Index (Li, 2008) and the length of annual reports (Loughran and McDonald, 2014), reduce the usefulness of text in these reports for investors (Lee, 2012; You and Zhang, 2009). Therefore, there has been an increasing demand for a new style of disclosure that provides relevant information that is easier to process for stakeholder groups. Integrated reporting (IR) has emerged as a potential response to this demand. This paper examines a key related research question: Do integrated reports mitigate the poor readability of text in annual reports?

IR is a global reporting framework conceptualised by the International Integrated Reporting Committee with the two-fold objective of improving communication to investors about a firm's value-creation process and promoting integrated thinking among firm managers. Integrated reports replace traditional annual reports[2] and provide a concise message about how a firm's strategy, governance, performance, and prospects lead to the creation of value over the short, medium, and long term. Not surprisingly, prior research on the consequences of IR adoption documents benefits related to the adoption; for example, Lee and Yeo (2016) and Barth *et al.* (2017) find a positive association between IR adoption and firm value.

However, there is limited empirical evidence on the role of the textual attributes of IR. Du Toit (2017) reports that integrated reports have higher linguistic complexity, where linguistic complexity is proxied by Flesch Reading Ease, the Flesch-Kincaid Formula, and the Gunning Fog Index. Caglio *et al.* (2020) document that the textual attributes of IR, such as readability, conciseness, and tone, are positively associated with economic benefits such as firm value, stock liquidity, and analyst forecast accuracy. In conjunction, these findings suggest that integrated reports use complex words and investors benefit from integrated reports that are linguistically less complex. However, what is unclear from the evolving literature on IR is the extent to which integrated reports successfully mitigate the linguistic complexity of reports and consequently become more informative to investors. In this paper, I attempt to fill this gap in the literature, focusing on the "connectivity" principle of IR.

Connectivity refers to combining all the pieces of a value-creation story cohesively in a single disclosure. It promotes integrated thinking, prompting managers to connect the various value drivers of a firm and communicate the same in a single report. If the principle of connectivity is followed in its true spirit, then the process of connecting the value drivers should improve the coherence of annual integrated reports. Improved text coherence is likely to improve the readability of reports by mitigating their linguistic complexity.

The linguistics literature identifies coherence as an essential element of writing since it

[2] In the remainder of the paper, the term "integrated reports" represents annual reports published after the mandate in South Africa. Thus, after 2010, annual reports and integrated reports convey the same meaning.

enables the text to convey its meaning (Bamberg, 1983; Grosz and Sidner, 1986). Coherence is defined as the ease with which a reader can understand a passage in a text given the knowledge obtained from reading the preceding passage. In other words, coherent text builds upon the preceding passage, thereby connecting all the passages of a text. Therefore, to the extent a piece of text with high coherence is easier to comprehend and serves the communication purpose (Wolf and Gibson, 2005), it is likely to reduce information processing costs for investors. Thus, I predict that the informativeness of integrated reports is higher for reports with higher coherence.

Next, I examine whether coherence mitigates the negative influence of traditional complexity on the informativeness of integrated reports. Given the central importance of coherence in linguistics, it should play a vital role in easing the processing of complex annual reports. In other words, a coherent integrated report is likely to mitigate the poor readability induced by the complex words used in integrated reports. On the basis of this, I argue that while integrated reports could have higher linguistic complexity, as reported by du Toit (2017) and Caglio *et al.* (2020), the coherence of such reports is likely to mitigate their complexity. Thus, coherence should make it easier for investors to process the information contained in these reports, consequently making them more informative. Since an integrated report combines complex pieces of information into a single document, the benefit of higher coherence should be higher for reports that have higher complexity. Therefore, I predict that the positive association between the coherence of integrated reports and the informativeness of these reports is stronger for reports that are more complex.

I begin the empirical analysis using a sample of 2,780 firm-year observations of firms listed on the Johannesburg Stock Exchange (JSE) and headquartered in South Africa during the period 2011 to 2019. I use latent semantic analysis (LSA) to compute the coherence of each annual report in my sample. LSA allows for calculating the cosine of the angle between two vectors, where each vector could represent a part of the text, such as a sentence or a paragraph. Details of this computation algorithm are provided in Appendix A.

I first examine whether the higher coherence of integrated reports is associated with the higher informativeness of these reports. To capture report informativeness, I measure the investor response to the release of these reports by computing the 3-day absolute cumulative abnormal returns (CAR) around the filing date of integrated reports by the firms in my sample. On the basis of the arguments presented earlier, I expect a positive association between the coherence of integrated reports and the informativeness of these reports. My findings support this prediction. Specifically, I find that the absolute 3-day (7-day) CAR of firms whose reports are more coherent are higher by approximately 1.5 (2.6) percentage points.

To further corroborate the above finding related to increased informativeness, I examine whether the coherence of integrated reports is associated with more information being impounded into stock prices. Market frictions, such as incomplete information or information

asymmetry, lead to stock price delay (Callen *et al.*, 2013; Hou and Moskowitz, 2005), which is a delayed adjustment of the stock price to new information. If the coherence of integrated reports makes these reports more informative, price discovery should also occur faster, leading to a reduction in stock price delay. Consistent with this prediction, I find that the stock price delay of firms with higher coherence reports is lower by approximately 6 percentage points.

The next cross-sectional test examines whether coherence mitigates the negative impact of the linguistic complexity[3] of reports. While prior studies have used report length and the Fog Index to capture the complexity of integrated reports (Caglio *et al.*, 2020; du Toit, 2017), these measures may not fully capture the textual attributes of integrated reports because IR emphasises coherence, which is a lesser studied attribute of disclosures in the literature. IR, through the principle of connectivity, is expected to combine various reports, such as sustainability reports and annual reports, into a single coherent disclosure that conveys a firm's value-creation story. Producing a report with high connectivity would require managers to integrate the various value drivers of the firm into a single coherent document. Thus, the coherence of integrated reports is likely to reduce the information processing costs for investors significantly. This benefit of coherence should be observed more for reports that have higher linguistic complexity and therefore poor readability. If this is true, then a coherent integrated report should mitigate the poor readability captured by the Fog Index. In other words, a coherent integrated report should be easier for investors to process despite the higher Fog Index and consequently should be more informative to investors. To test this prediction, I examine whether the association between report informativeness and coherence varies with the Fog Index of integrated reports. I find that the relationship between CAR (stock price delay) and the coherence of reports is more positive (more negative) for firms with a higher Fog Index.

This study contributes to several streams of the literature. First, it adds to the evolving textual literature on IR and shows that the coherence of integrated reports is crucial to the usefulness of IR to investors. In particular, this study complements the works of Caglio *et al.* (2020) and du Toit (2017) and adds a new dimension to the text-based research on IR. Caglio *et al.* (2020) document that integrated reports are useful for an investor when these reports have lower linguistic complexity, where complexity is measured using traditional measures such as the Fog Index. This paper builds upon their work and proposes that integrated reports could be beneficial to investors even when the linguistic complexity is high, provided that these reports are written coherently. Thus, this paper adds to the existing literature on IR by moving one step closer to understanding the textual attributes of integrated reports.

Second, this paper aids the regulators in their endeavour to achieve improved readability of corporate disclosures by highlighting the role of coherence in mitigating the effect of

[3] The terms "linguistic complexity" and "complexity" are used interchangeably in this paper.

complex words in a document. The finding that coherence in a report gains more importance when the words used in the report are complex sheds light on the future path that firms could follow to improve transparency in their disclosures without necessarily reducing the usage of complex words. The findings in this study could be easily extended to other markets owing to the ease of computation of the coherence measure and the large sample availability of annual reports.

Finally, this study adds to the literature that examines the usefulness of IR to investors. Prior literature on IR shows the economic consequences of IR adoption, such as an improvement in analyst forecast accuracy, a reduction in the cost of capital, and an increase in investment efficiency (Baboukardos and Rimmel, 2016; Bernardi and Stark, 2018; Zhou *et al.*, 2017). This study adds to the literature by documenting a potential channel through which some of these economic benefits could accrue to the firm.

II. Background and Hypotheses Development

2.1 Integrated Reporting—Institutional Background

The global financial crisis of 2007 was an inflection point for reporting frameworks. Investors and creditors demanded clear and relevant information regarding value creation, risk management, and external factors that influence business. With the demand for a new style of disclosure increasing with time, there was a need to create a globally accepted framework that results in comprehensive communications by an organisation about value creation over time. Today, this new style of reporting framework is known as IR. IR facilitates the presentation of all the value drivers of a firm in a single report. It also provides a synergy between those value drivers so that investors can understand the value-creation story of a firm in conjunction with the risk embedded in the firm and its risk management practices.

In 2009, the Prince of Wales chaired a meeting of various stakeholder bodies, such as investors, companies, regulators, and standard setters, to establish the International Integrated Reporting Committee, a body to supervise the creation of a globally accepted IR framework. This body was officially created in August 2010 and was renamed the International Integrated Reporting Council (IIRC) in November 2011.

The IIRC, in its International Integrated Reporting Framework, defines an integrated report as a concise communication about how an organisation's strategy, governance, performance, and prospects, in the context of its external environment, lead to the creation of value in the short, medium and long term (IIRC, 2021). With the IIRC starting a pilot programme to develop an IR framework that included 90 businesses, South Africa started a transition to IR. In February 2010, the King III Codes of Governance were made a mandatory part of listing requirements in the JSE for South African firms. One of these requirements was to prepare an integrated report on a 'comply or explain' basis. To assist firms, a voluntary,

not-for-profit organisation called the Integrated Reporting Committee of South Africa (IRC) was formed in 2010. The IRC created a framework to ease the process of transitioning from traditional reporting to IR.

One of the key elements of an integrated report is the *connectivity* or *integration* of information. Connectivity promotes integrated thinking among the managers of a firm, leading to better strategic decision-making, more connected departments within the firm, and improved internal processes. IR combines the various value drivers of a firm in a single report. For example, an integrated report describes an organisation's business model and its connection with the six types of capital identified by the IIRC (financial, human, intellectual, manufactured, social, and relationship capital). The report communicates to investors the extent to which the business depends on these types of capital, thereby highlighting the potential risks and opportunities that the organisation faces.

This concept of integrated thinking differentiates IR from traditional reporting. For example, the Management Discussion and Analysis (MD&A) section of an annual report provides information on firm performance in the prior year and future projections. It also discusses key trends and risks in the business. The information, however, is not presented in a connected fashion. IR, on the other hand, documents all the resources, or capitals, of a firm and how these resources are linked to the firm's strategy. Traditional reporting occurs in silos, but IR connects these silos to present a holistic picture of the organisation.

The annual integrated report of 2015 prepared by Kumba Iron Ore Ltd, a major South African supplier of iron ore to the global steel industry, is an example of how the connectivity of information is the core of an integrated report. Kumba's report integrates strategy, the business model, the operating context, risks and opportunities, and governance. The six types of capital are introduced early in the report with key inputs, and the outcomes of each capital are clearly specified. The actions needed to achieve these outcomes are also detailed. The report uses a diagrammatic representation of the business model to achieve this connectivity.

The IIRC believes that an integrated report should explain the reporting entity's interrelated financial, environmental, social, and corporate governance information. At the same time, it should be presented in a clear, concise, consistent, and comparable manner. To help organisations transition to IR from traditional reporting, the IIRC proposes a set of guiding principles. These principles assist firms to prepare integrated reports that achieve the objectives of integrated thinking and effective communication to investors. According to the IIRC, an integrated report should report on the following dimensions: (1) organisational overview, (2) governance mechanisms, (3) business model overview, (4) risks that a firm faces and existing and future opportunities, (5) strategy formulation and resource allocation mechanism and structure, (6) dimensions of organisational performance and its metrics, and (7) future orientation and outlook. These seven points are the pillars of the official IR framework issued by IIRC in 2013.

2.2 Coherence of Integrated Reports

The objectives of IR aim at the unification of all pieces of relevant information into a single report. An integrated report is expected to provide financial and non-financial information in a 'cohesive' fashion (Caglio *et al.*, 2020). Thus, IR is expected to improve the readability of reports through the principle of connectivity by connecting all the pieces of a value-creation story in a cohesive manner. The principle of connectivity is closely linked to the linguistic notion of coherence. Coherence is a vital element of writing because it enables the text to convey its meaning (Bamberg, 1983; Grosz and Sidner, 1986). It is defined as the ease with which a reader can understand a passage in a text given the knowledge obtained from reading the passage immediately preceding the current one. In other words, a coherent text builds upon the preceding passage, thereby connecting all the passages of a text. Therefore, a piece of text that is highly coherent is easier to comprehend and serves the communication purpose (Wolf and Gibson, 2005). Thus, if the IIRC guidelines on connectivity of information are followed in their true spirit, the coherence of annual integrated reports is likely to reduce the information processing costs for investors. Thus, integrated reports with higher coherence should prove to be more informative to investors. This discussion leads to the following hypothesis:

H1: The informativeness of an integrated report is positively associated with its coherence.

2.3 Coherence of Integrated Reports and Stock Price Delay

Lawrence (2013) documents that simple disclosures attract investors. In other words, disclosures with higher linguistic complexity lead investors to neglect stocks, causing stock price delay (Callen *et al.*, 2013; Hou and Moskowitz, 2005). Thus, if coherence improves the readability of integrated reports, the stock price delay should be reduced. Callen *et al.* (2013) document that when the market-wide component of information is held constant, the quality of the pre-existing information set influences the speed at which the stock price adjusts to the arrival of news, also known as stock price delay. They build upon the work of Hou and Moskowitz (2005), who document that market imperfections such as information symmetry lead to stock price delays. Callen *et al.* (2013) show that poor accounting quality renders the investors' pre-existing information set inferior, leading to a slower adjustment of stock prices to news. Based on their arguments, if the pre-existing information set of investors can be improved, stock price delay should be reduced. IR, through a reduction in information processing costs, could potentially achieve this. Therefore, I predict that the stock prices of firms with coherent reports will adjust faster to newly arriving information. This discussion leads to the following hypothesis:

H2: Stock price delay is negatively associated with the coherence of integrated reports.

2.4 Linguistic Complexity, Coherence, and Informativeness

Prior literature in accounting and finance captures the linguistic complexity of disclosures using several textual attributes, such as the Fog Index (Li, 2008), report length (Loughran and McDonald, 2011), and file size (Loughran and McDonald, 2014). There is ample evidence from prior studies that these measures are strongly associated with capital market outcomes. Cazier and Pfeiffer (2015) show that price discovery is slower for firms that produce long annual reports with excessive boilerplate. Brown and Tucker (2011) find that when the MD&A section borrows text from the previous year and contains relatively less new information, it evokes a low response from the market at the time of filing. Dyer *et al.* (2017) find that boilerplate text in annual reports is positively associated with measures of information asymmetry, such as liquidity, analyst following, and institutional ownership.

In the case of IR, there is little documentation of its textual attributes. Du Toit (2017) provides small-sample evidence of an increase in the linguistic complexity of integrated reports, where linguistic complexity is captured using traditional measures such as Flesch Reading Ease, the Flesch-Kincaid Formula, and the Gunning Fog Index. Caglio *et al.* (2020) suggest that the linguistic complexity of integrated reports should be measured using traditional measures, such as length and the Fog Index, and find that IR is beneficial to investors when integrated reports are concise and have a lower Fog Index.

I complement the findings of Caglio *et al.* (2020) and suggest that IR could be beneficial to investors even when these reports have a higher Fog Index. IR is expected to reduce the linguistic complexity of reports through the principle of connectivity by connecting all the pieces of the value-creation story in a cohesive manner. The connectivity principle ensures that the relationships among key elements included in the report are explicitly and clearly presented. Investors observe the connectivity of a report by reading it. Producing a report with high connectivity would require managers to integrate the various value drivers of the firm into a single coherent document. Thus, the coherence of integrated reports is likely to reduce the information processing costs for investors significantly. This benefit of coherence should be observed more for reports that have higher linguistic complexity, and therefore poor readability, to begin with. If a report uses simple words and already has a high readability, the scope to further improve the readability via coherence is not very significant. However, if a report uses complex words and therefore has poor readability, coherence could play a pivotal role in reducing the complexity of the document by connecting the complex words in such a way that the readability of the document improves significantly. Therefore, I expect coherence to be more beneficial for reports with higher linguistic complexity or poor readability. This discussion leads to the following hypotheses:

H3A: The relationship between coherence and informativeness is more positive for firms with a higher Fog Index.

H3B: The relationship between coherence and stock price delay is more negative for firms with a higher Fog Index.

III. Research Framework and Empirical Results

3.1 Data and Sample

I collect the data for testing the hypotheses from various sources. I obtain firm-year observations for firms domiciled in South Africa and listed on the JSE for the period 2011 to 2019 from Compustat Global. This period corresponds to the years after IR was mandated, when firms had to either follow IR or explain the reason if they chose not to follow IR. I combine these data with daily stock price data from the Compustat Securities file to compute stock price delay. To create a coherence score based on LSA, I obtain annual reports in PDF format for all firms in my sample from 2011 to 2019 from S&P Capital IQ. Annual reports are not available for 131 firm-years during this period. Of the available reports, I retain only those that could reasonably be identified as integrated reports. To ensure this, I run a search algorithm for the term "integrated report" in each report and only retain those that return a minimum of five[4] occurrences of this term. This algorithm leads to a rejection of 378 downloaded reports. Thus, a total of 509 observations are excluded from the Compustat universe on the basis of these criteria. Table 1 presents details of the sample construction.

Table 1 Sample Selection

	DROPPED	SAMPLE SIZE
Compustat Global Data for South African firms 2011–2019		2,780
Drop missing observations for control variables:		
Drop observations with no annual integrated report available	(509)	2,271
Drop missing control variables from Abret Regression	(221)	2,050
Drop missing control variables from Delay Regression	(345)	1,705
FIRM-YEAR SAMPLE		1,705

3.2 Measuring Coherence

Investors observe the connectivity of a report by reading it. If the manager has connected

[4] The choice of five occurrences is not based on prior literature but merely on an empirical observation of my sample. For example, the 2017 integrated report of Kumba Iron has 15 mentions of the term "integrated report", excluding the occurrences in the footer of each page. On the other hand, some firms mention this term just to explain the reason for not complying with IR. Keeping five occurrences as a threshold ensures that the firm is following IR.

various value-creating elements to present the value-creation story to investors, this should be reflected in the text of the report. Connectivity relates to the concept of coherence in psycholinguistics. Pinker (2014) proposes that a coherent text conveys the idea efficiently, while if the text is not coherent, the reader disregards the information it presents. To capture this coherence, I follow prior work in the linguistics literature. Linguistics theory defines coherence as the characteristic of textual discourse that allows a reader to "move easily from one sentence to the next and read the paragraph as an integrated whole" (Bamberg, 1983, p. 417). This definition of coherence is effectively captured in the *coherence* function of the R software package. The coherence function uses LSA to create a measure of the dependence of a given paragraph in the text on the preceding paragraph. This measure of dependence conveys the coherence of the text.

Foltz *et al.* (1998) propose a simple algorithm to capture the coherence of a text using LSA. They posit that coherence can be computed by comparing one text unit to an adjoining text unit and then measuring the degree to which the two are semantically related. The unit of text could be a sentence or a paragraph. Crossley and McNamara (2011) suggest that this notion of textual coherence is similar to the concept of the "givenness" of text. Givenness captures the proportion of given information to new information in a text. In other words, givenness represents the amount of information that can be recovered from the preceding discourse. Not surprisingly, the linguistics literature documents a strong positive association between the givenness of text and coherence. Consequently, givenness has been widely used as a proxy for coherence (Crossley and McNamara, 2011). Therefore, I use the givenness of text as the measure of coherence in this paper, where I capture givenness using the semantic relatedness of adjoining units of text.

To measure semantic relatedness, prior literature has proposed examining the cosines between the vectors of two units of text (Higgins and Burstein, 2007; Foltz *et al.*, 1998; Günther *et al.*, 2015). This cosine measure is labelled as the coherence between the two units of text. The cosine between two adjoining sentences is known as local coherence, and the average value of the cosines for sentences of two adjoining paragraphs is known as global coherence. The *LSAfun* package in R has an inbuilt function known as 'coherence' that computes both coherence measures. For this study, I focus on global coherence because coherence between paragraphs is likely to be more informative to the readers, and label global coherence as coherence, denoted by *coherence*. I scale the variable to allow it to vary between 0 and 1. Further details on computing coherence are provided in Appendix A.

3.3 Measuring Stock Price Delay

Following Hou and Moskowitz (2005) and Callen *et al.* (2013), I calculate the average delay with which information is absorbed into stock prices by first regressing stock returns for each firm on contemporaneous market returns and four lagged market returns as follows:

$$r_{i,t} = \alpha_i + \beta_i R_{m,t} + \sum_{n=1}^{4} \delta_{i,n} R_{m,t-n} + \epsilon_{i,t}, \quad (1)$$

where $r_{i,t}$ is the return on stock i and $R_{m,t}$ is the market return in week t. The underlying concept of a stock price delay is the lagged stock price response to market news. If the stock price response to new information is delayed, returns from the prior period will have explanatory power for contemporaneous stock returns. In such a case, $\delta_{i,n}$ could be non-zero. This is an unrestricted regression. In the case of no stock price delay, all $\delta_{i,n}$ will be equal to zero. This is a restricted regression. Stock price delay is defined as

$$Delay = 1 - \left(\frac{R^2_{restricted}}{R^2_{unrestricted}}\right) \quad (2)$$

When the delay is larger, lagged returns explain some variance in contemporaneous returns. Model (1) is estimated using weekly returns from $July_{t-1}$ to $June_t$ to calculate $Delay_t$. The model uses market returns, or systematic news, as the stimulus to which stock i responds. In this manner, newly arriving market-wide information is held constant.

Delay computed at the individual stock level may induce estimation error. To mitigate this, I estimate delay at the portfolio level and use it for my main specification. I first sort the firms into deciles on the basis of their size and then sort them into deciles on the basis of the stock-level delay measure computed earlier. I then re-compute the delay on the basis of portfolio returns.

3.4 Empirical Framework and Results

3.4.1 Test for informativeness

Hypothesis H1 predicts that the informativeness of integrated reports is positively associated with the coherence of these reports. Following prior literature (see, for example, Merkley (2013)), I capture the informativeness of reports by measuring investor response to their release. Following prior studies (Bushee *et al.*, 2010; Rogers and Van Buskirk, 2009), I examine investor response to the issuance of integrated reports by using the absolute market-adjusted return in a short window around the filing of the integrated report. The model used is as follows:

$$(Abret1 \ (or \ Abret3)) = \beta_0 + \beta_1 \ HighCoherence + \Sigma\beta_i \boldsymbol{Controls} + Firm \ Fixed \ Effects + \varepsilon \quad (3)$$

In equation (3), *Abret1 (Abret3)* is the absolute market-adjusted CAR for a firm in the [-1,+1] ([-3,+3]) window around the filing of its annual report; *HighCoherence* is a dummy variable that takes the value of 1 if a firm's annual report has above median coherence in a given year; and $\boldsymbol{Controls}$ is a vector of control variables. Following Merkley (2013), I control for firm characteristics as in equation (3), change in earnings (*adjROA)*, and the number of analysts following the firm (*nanalyst).* Also, following prior studies (Lang and Stice-Lawrence, 2015; Li, 2008), I control for commonly known determinants of linguistic

complexity, such as firm size (*mktval*), age of the firm (*age*), leverage (*lev*), market-to-book ratio (*mtb*), earnings surprise (*earnsurp*), and an indicator for loss making firms (*loss*), along with the Fog Index (*fog*) and length of annual reports (*length*). I additionally control for the most recent earnings announcement ($Abret_{EA}$). The definitions of all variables are presented in Table 2. I winsorise all continuous variables at the 1% and 99% levels. Summary statistics of these variables are shown in Table 3 Panel A, and the correlation matrix is presented in Table 3 Panel B.

Table 2 Variable Descriptions

VARIABLE	NOTATION	DEFINITION/MEASUREMENT	LEVEL (FIRM / FIRM-YEAR)
CAR[-1,1]	*Abret1*	Cumulative abnormal return 3 days around the filing of annual report	Firm-Year
CAR[-3,3]	*Abret3*	Cumulative abnormal return 5 days around the filing of annual report	Firm-Year
Portfolio Delay	*Delay*	Delay computed at portfolio level based on the procedure outlined in Callen *et al.* (2013)	Firm-Year
Coherence	*Coherence*	Coherence measure described in Appendix A	Firm-Year
High Coherence	*HighCoherence*	A dummy variable that takes the value of 1 if *coherence* is above its median value in a given year and 0 otherwise	Firm-Year
Market Value of Equity	*Mktval*	Stock price at the end of year * total shares outstanding at the end of the year	Firm-Year
Adjusted ROA	*adjROA*	Change in ROA over the previous year. ROA is measured as net income scaled by lagged total assets	Firm-Year
Number of Analysts	*nanalyst*	Natural log of number of analysts from IBES following the firm	Firm-Year
CAR around earnings announcement	*abret_EA*	3 day or 5 day CAR around the recent most earnings announcement of a firm	Firm-Year
Loss Frequency	*Lossfreq*	Number of times a firm reports loss in last four years	Firm-Year
Share Turnover	*Shturn*	Trading volume of a stock	Firm-Year
Growth Opportunities	*Mtb*	*mktval* / book value of a firm	Firm-Year
Fog Index	*Fog*	Gunning Fog Index	Firm-Year
High Fog Index	*HighFog*	A dummy variable that takes the value of 1 if *fog* is above its median value in a given year and 0 otherwise	Firm-Year
Length	*length*	Natural log of the number of words in a 10-K, after cleaning the text	Firm-Year
Loss Indicator	*loss*	Dummy variable that equals 1 if net income < 0 and 0 otherwise	Firm-Year
Earnings Surprise	*earnsurp*	Change in earnings (eps) over last year, scaled by closing market price of stock	Firm-Year
Firm Age	*age*	Natural log of the number of years in COMPUSTAT since IPO	Firm-Year

Table 3 Descriptive Statistics

Table 3 presents the summary statistics of the key variables used to test the main hypotheses. The sample spans 9 years from 2011 to 2019 (N=1,705). All continuous variable are winsorised at the 1% and 99% levels. All variables are described in Table 2. Panel A reports the summary statistics for all variables. Panel B presents the Spearman (above diagonal) and Pearson (below diagonal) correlation coefficients for all variables. Values in bold indicate statistical significance atthe 1% level or better.

Panel A Summary Statistics

VARIABLES	(1) N	(2) Mean	(3) Sd	(4) P25	(5) Median	(6) P75
DEPENDENT VARIABLES						
Abret1	1,705	0.046	0.016	0.032	0.032	0.046
Abret3	1,705	0.077	0.032	0.047	0.047	0.089
Delay	1,705	0.062	0.081	0.014	0.014	0.039
IR VARIABLES						
Coherence	1,705	0.451	0.151	0.129	0.415	0.655
HighCoherence	1,705	0.500	0.500			
CONTROL VARIABLES						
Mktval	1,705	6.755	2.191	5.154	5.154	6.768
adjROA	1,705	-0.003	0.079	-0.004	-0.004	-0.004
nanalyst	1,705	6.004	2.603	5.000	5.000	6.000
abret_EA	1,705	0.025	0.026	0.005	0.026	0.036
Lossfreq	1,705	0.211	0.288			
Shturn	1,705	6.980	0.905	6.420	6.420	7.026
Mtb	1,705	1.548	2.404	0.887	0.887	1.146
Age	1,705	2.192	0.523	1.792	2.198	2.639
Loss	1,705	0.263	0.441			
earnsurp	1,705	-0.003	0.001	-0.002	0.000	0.002
TEXTUAL COMPLEXITY VARIABLES						
Fog	1,705	24.101	2.390	21.854	23.131	25.637
HighFog	1,705	0.500	0.500			
Length	1,705	10.461	0.411	9.714	10.153	10.615

Panel B Correlation Matrix

	Abret1	*Abret3*	*Delay*	*coherence*	*mktval*	*adjROA*	*nanalyst*	*Abret_EA*	*lossfreq*	*shturn*	*mtb*	*loss*	*Delay*	*fog*	*length*	*age*	*earnsurp*
Abret1		**0.346**	**-0.06**	**0.09**	**-0.196**	**0.016**	0.012	-0.038	0.251	0.018	-0.004	0.263	-0.056	-0.021	0.019	0.06	-0.013
Abret3	**0.298**		**-0.1**	**0.128**	**-0.156**	**0.082**	**0.056**	**-0.066**	0.156	-0.003	0.011	0.087	-0.102	-0.033	0.023	0.037	-0.022
Delay	**-0.108**	**-0.147**		**-0.052**	**-0.023**	**0.013**	-0.232	0.044	-0.016	-0.292	-0.054	0.028	1	0.076	-0.027	-0.015	0.057
coherence	**0.051**	**0.078**	**-0.1**		**-0.207**	**0.003**	**0.016**	-0.067	0.116	0.005	-0.005	0.12	-0.052	-0.019	-0.057	0.197	-0.016
mktval	**-0.022**	**-0.15**	**-0.15**	**-0.203**		**0.015**	**0.259**	**0.055**	**-0.378**	**0.111**	**0.233**	**-0.367**	**-0.023**	-0.105	0.282	0.121	0.118
adjROA	**0.043**	**0.048**	**-0.03**	**0.022**	0.044		-0.01	0.028	0.012	-0.031	0.107	-0.201	0.013	-0.002	0.012	0.006	0.231
nanalyst	**0.143**	**0.103**	**-0.48**	**0.114**	**0.144**	**0.064**		**-0.064**	**-0.017**	**0.281**	0.049	-0.03	-0.232	-0.029	0.117	0.084	0.077
abret_EA	**0.004**	**-0.075**	0.026	-0.027	0.202	0.025	-0.061		-0.096	-0.004	-0.024	-0.097	0.044	-0.042	-0.04	0.08	0.013
lossfreq	**0.095**	**0.141**	**0.018**	**0.114**	**-0.377**	**-0.054**	-0.002	-0.101		-0.004	0.032	0.699	-0.016	0.021	0.012	-0.093	-0.012
shturn	**0.014**	**-0.017**	**-0.2**	**-0.005**	**0.131**	**-0.033**	0.174	0.036	-0.011		0.015	-0.03	-0.292	-0.097	0.122	-0.032	-0.007
mtb	**-0.053**	**-0.11**	**-0.09**	**-0.125**	**0.533**	**0.104**	0.071	0.042	-0.23	0.072		-0.021	-0.054	0.007	0.019	0.046	0.046
loss	**0.1**	**0.094**	**0.025**	**0.128**	**-0.361**	**-0.23**	**-0.004**	**-0.111**	**0.684**	-0.029	-0.277		0.028	0.097	0.029	-0.028	-0.199
Delay	**-0.108**	**-0.147**	**1**	**-0.101**	**-0.153**	**-0.028**	**-0.483**	**0.026**	0.018	-0.2	-0.085	0.025		0.076	-0.027	-0.015	0.057
fog	**-0.007**	**-0.053**	**0.134**	**-0.051**	**-0.193**	**-0.024**	**-0.072**	**-0.131**	0.035	-0.085	-0.122	0.11	0.134		-0.132	-0.042	-0.079
length	**0.072**	**0.019**	**-0.06**	**-0.093**	0.306	-0.026	0.059	0.008	0.062	0.126	0.079	0.055	-0.062	-0.154		0.021	0.059
age	**0.118**	**-0.021**	**-0.2**	**0.142**	**0.1**	**0.046**	**0.295**	**0.163**	**-0.093**	-0.018	-0.047	-0.004	-0.203	-0.08	0.011		-0.009
earnsurp	**0.006**	**0.005**	**-0.05**	**0.001**	**0.203**	0.454	0.056	0.04	-0.023	0.041	0.214	-0.331	-0.054	-0.087	0.009	0.05	

I include firm fixed effects and year fixed effects to control for the unobserved characteristics along these dimensions. Standard errors are clustered at year and industry level, where industry is defined using the Fama-French (12) classification. Hypothesis H1 predicts a positive and significant coefficient β_1 on *HighCoherence*. The results from this test are presented in Table 4.

Table 4 Investor Response to Filing of Integrated Reports—Role of Coherence

Table 4 reports the results from estimating equation (4) using OLS. The sample is comprised of 1,705 firm-years spanning the period 2011 to 2019. T statistics, based on standard errors clustered by year and industry (Fama French 12), are included in parentheses. Two-tailed p-values are indicated: *** p < 0.01, ** p < 0.05, * p < 0.10. All variables are described in Table 2. Column 1 (2) reports the results for 3-day (7-day) CAR.

	(1)	(2)
VARIABLES	*Abret1*	*Abret3*
HighCoherence	**0.0150*** **	**0.0260*** **
	(3.0386)	**(2.7532)**
CONTROL VARIABLES		
mtb	-0.0004**	-0.0037***
	(-2.3610)	(-3.2220)
mktval	-0.0050**	-0.0177*
	(-2.1066)	(-1.7353)
adjROA	-0.1043*	-0.0994***
	(-1.9054)	(-2.7990)
nanalyst	0.0013	0.0019
	(0.7241)	(0.9318)
age	-0.0011**	-0.0201*
	(-2.2791)	(-1.7801)
earnsurp	0.0051**	0.0061**
	(2.1922)	(2.2133)
loss	-0.0041**	-0.0051*
	(-2.1331)	(-1.7018)
fog	-0.0035**	-0.0011**
	(-2.2146)	(-2.1345)
length	-0.0016	-0.0006
	(-0.9509)	(-0.1783)
Abret_EA	0.1731*	0.1090*
	(1.6934)	(1.7657)
Constant	0.3439*	0.2359*
	(1.7339)	(1.8738)
Observations	1,705	1,705
R-squared	0.757	0.496
FIRM FE	YES	YES
YEAR FE	YES	YES

In column (1) of Table 4, the dependent variable is *Abret1*. The coefficient β_1 is +0.0150, significant at the 1% level, suggesting that the 3-day CAR are higher by 1.5 percentage points for firms with higher coherence. Column (2) repeats the test with the dependent variable as *Abret3*. The findings in column (2) are qualitatively similar to those in column (1). Taken

together, I document evidence of the role of the coherence of annual reports in increasing the investor response to the issuance of integrated reports, thus finding support for hypothesis H1.

3.4.2 Tests for stock price delay

Hypothesis H2 predicts that if the coherence of reports leads to a higher response from investors, the information contained in these reports should be impounded faster into stock prices, thereby reducing stock price delay. I examine this hypothesis using the following model:

$$\begin{aligned} Delay \;=\; & \beta_0 + \beta_1 HighCoherence + \beta_2 mktval + \beta_3 shturn \\ & + \beta_4 nanalyst + \beta_5 lossfreq + \beta_6 Fog + \beta_7 Length \\ & + Firm\ \ Fixed\ \ Effects + Year\ \ Fixed\ \ Effects + \varepsilon \end{aligned} \tag{4}$$

Table 5 Stock Price Delay—Role of Coherence

Table 5 reports the results from estimating equation (5) using OLS. The sample is comprised of 1,705 firm-years spanning the period 2011 to 2019. T statistics, based on standard errors clustered by year and industry (Fama French 12), are included in parentheses. Two-tailed p-values are indicated: *** p < 0.01, ** p < 0.05, * p < 0.10. All variables are described in Table 2.

	(1)	(2)
VARIABLES	*Delay*	*Delay*
HighCoherence	**-0.0742****	**-0.0612****
	(-2.1254)	**(-2.2171)**
CONTROL VARIABLES		
Mktval	-0.0074*	-0.0056**
	(-1.7137)	(-2.1028)
shturn	-0.0123**	-0.0117*
	(-2.2061)	(-1.9156)
nanalyst	-0.0310	-0.0170
	(-0.5614)	(-0.7473)
lossfreq	0.0143**	0.0157**
	(2.1343)	(2.1557)
fog		0.0121**
		(2.1813)
length		0.0246
		(1.3615)
Constant	-0.0413	-0.0454
	(-0.5113)	(-0.3063)
Observations	1,705	1,705
R-squared	0.565	0.613
FIRM FE	YES	YES
YEAR FE	YES	YES

In equation (4), *Delay* is the portfolio-level delay measure. Following Callen *et al.* (2013) and Hou and Moskowitz (2005), I control for the size of the firm (*mktval*) and the liquidity of a firm using share turnover as a proxy (*Shturn*). I also control for the firm's information environment using the natural log of the number of analysts following a firm (*nanalyst*).

Finally, I include *lossfreq*, which is the number of times a firm has reported a loss in the past 3 years. On the basis of hypothesis H2, I predict a negative coefficient β_1 on *HighCoherence*.

Table 5 column (1) presents the results from this test. The coefficient on *HighCoherence* is -0.0742, significant at the 5% level. This suggests that, on average, the stock price delay is lower by approximately 7.4 percentage points for integrated reports with higher coherence. These findings indicate that IR allows investors to focus their attention on neglected stocks, thus reducing stock price delays for these stocks.

In column (2) of Table 5, I also control for textual complexity measures such as the Fog Index and length of the report. The coefficient β_1 is -0.0612, significant at the 5% level, suggesting that the stock price delay is lower by approximately 6.1 percentage points for integrated reports with high coherence, even after controlling for the linguistic complexity of these reports.

3.4.3 Tests for role of coherence in mitigating linguistic complexity

The next set of hypotheses examines whether the coherence of integrated reports mitigates the negative influence of linguistic complexity on investor response and stock price delay. I argue that coherence gains more importance in conveying the message of a text when the text has higher linguistic complexity. If the text uses complex words, coherence can mitigate the impact of these words on the reader by making the overall text easier to read. Hypothesis H3A tests for the role of coherence in mitigating linguistic complexity on investor response. To test this hypothesis, I use the following model:

$$\begin{aligned} Abret = \ & \beta_0 + \beta_1 \ HighCoherence * HighFog + \beta_2 \ HighCoherence \\ & + \ \beta_3 \ HighFog + \Sigma\beta_i \boldsymbol{Controls} + Firm \ \ Fixed \ \ Effects \\ & + Year \ \ Fixed \ \ Effectts + \varepsilon \end{aligned} \tag{5}$$

In equation (5), *HighFog* is a dummy variable that takes the value of 1 if the annual report of a firm has an above median Fog Index in a given year. If coherence helps investors understand a complex text better, investor response should be higher for firms with higher complexity and higher coherence. The coefficient β_1 captures the incremental influence of the high coherence of reports on investor reaction for higher Fog Index reports. If coherence is more useful for investors when annual reports have higher traditional complexity than when these reports have lower traditional complexity, β_1 should be positive and significant. The results from this test are presented in Table 6.

Column (1) of Table 6 presents the results for 3-day CAR. The coefficient β_1 is +0.0078, significant at the 1% level. This suggests that investor reaction is higher by 78 basis points when complex reports have higher coherence. Thus, the traditional complexity of integrated reports is mitigated by their high coherence, leading to higher information dissemination, as suggested by an increase in CAR. Column (2) of Table 6 documents the results for the 7-day CAR. The results are similar to those in column (1). This suggests that investor response

deteriorates when a report is complex to read and has poor coherence in the cross-section of firms.

Table 6 Role of Coherence in Mitigating the Influence of the Fog Index on Investor Response

Table 6 reports the results from estimating equation (6) using OLS. The sample is comprised of 1,705 firm-years spanning the period 2011 to 2019. T statistics, based on standard errors clustered by year and industry (Fama French 12), are included in parentheses. Two-tailed p-values are indicated: *** p < 0.01, ** p < 0.05, * p < 0.10. All variables are described in Table 2. Column 1 (2) reports the results for 3-day (7-day) CAR. The control variables used are the same as those in Table 4.

	(1)	(2)
VARIABLES	*Abret1*	*Abret3*
HighCoherence*HighFog	**0.0078*** **	**0.0089** **
	(4.1551)	**(2.2158)**
HighCoherence	0.0086**	0.0123***
	(2.1582)	(5.9566)
HighFog	-0.0047***	-0.0051***
	(-5.6856)	(-5.3749)
CONTROL VARIABLES	YES	YES
Observations	1,705	1,705
R-squared	0.721	0.883
FIRM FE	YES	YES
YEAR FE	YES	YES

Table 7 Role of Coherence in Mitigating the Influence of the Fog Index on Stock Price Delay

Table 7 reports the results from estimating equation (7) using OLS. The sample is comprised of 1,705 firm-years spanning the period 2011 to 2019. T statistics, based on standard errors clustered by year and industry (Fama French 12), are included in parentheses. Two-tailed p-values are indicated: *** p < 0.01, ** p < 0.05, * p < 0.10. All variables are described in Table 2. Column 1 (2) reports the results for 3-day (7-day) CAR. All variables are described in Table 2. The control variables used are the same as those in Table 5.

	(1)
VARIABLES	*Delay*
HighCoherence*HighFog	**-0.0534*** **
	(-3.9167)
HighCoherence	-0.0786**
	(2.2385)
HighFog	0.0347***
	(5.5121)
CONTROL VARIABLES	YES
Observations	1,705
R-squared	0.586
FIRM FE	YES
YEAR FE	YES

Using a similar argument, I test hypothesis H3B using the following model:

$$Delay = \beta_0 + \beta_1\ HighCoherence * HighFog + \beta_2\ HighCoherence + \beta_3\ HighFog + \beta_4 mktval + \beta_5 shturn + \beta_6 nanalyst + \beta_7 lossfreq + \beta_8 Fog + \beta_9 Length + Firm\ Fixed\ Effects + \varepsilon \quad (6)$$

Hypothesis H3B posits that reports that more complex to read would lead to a lower stock price delay if they have higher coherence. On the other hand, complex reports with low coherence would increase the stock price delay. Thus, the coefficient β_1 should be negative. The results from this test are presented in Table 7.

In Table 7, the coefficient β_1 is -0.0534, which is statistically significant at the 1% level. This finding suggests that reports with higher complexity lead to a lower stock price delay if they are coherent to read. Thus, coherence mitigates the complexity of annual reports, leading to the faster adjustment of stock price to firm-level information.

IV. Conclusion

Prior studies document the diminishing response of investors to the filing of annual reports. These studies support the regulatory concern of a reduction in the ability of annual reports to convey relevant information to investors. In response, IR attempts to mitigate the poor readability of annual reports by connecting all pieces of relevant information into a single report. The endeavour is to supply all information to investors through a single document that is coherent to read. I examine whether coherence mitigates the traditional complexity of disclosures by making them more informative to investors. I find that the 3-day (and 7-day) absolute CAR around the filing of integrated reports are higher for firms with a higher coherence in their integrated reports, suggesting that coherent reports are more informative to investors. I further examine whether coherence gains more importance when reports are more complex to read, as captured by the Fog Index. I find evidence to support that coherence mitigates the negative impact of the Fog Index on investors and is an important tool to understand reports when they have a high Fog Index. These findings highlight the role played by the coherence of information contained in integrated reports in increasing the ease of information processing by investors.

On the basis of these findings, this study lends support to the usefulness of IR as a response to mitigate the poor readability of annual reports. IR was mandated in South Africa with the objective of reducing the complexity of disclosures by making reports more coherent, allowing investors to access all relevant information in one place in a cohesive and integrated fashion. This study finds evidence that IR succeeds in achieving these objectives.

However, there is further scope for research to examine the role of textual attributes in the context of IR. A recent study by Siano and Wysocki (2018) highlights the role of financial numbers embedded in the text of disclosures in determining the complexity of these disclosures. Future research could examine whether financial information in the textual

portion of integrated reports impacts the coherence of these reports and whether the coherence mitigates the complexity induced by the financial numbers.

References

Baboukardos, D. and Rimmel, G. (2016), 'Value Relevance of Accounting Information Under an Integrated Reporting Approach: A Research Note', *Journal of Accounting and Public Policy* 35 (4): 437–452.

Bamberg, B. (1983), 'What Makes a Text Coherent?', *College Composition and Communication* 34 (4): 417–429.

Barth, M. E., Cahan, S. F., Chen, L., and Venter, E. R. (2017), 'The economic consequences associated with integrated report quality: Capital market and real effects', *Accounting, Organizations and Society* 62: 43–64.

Bernardi, C. and Stark, A. W. (2018), 'Environmental, Social and Governance Disclosure, Integrated Reporting, and the Accuracy of Analyst Forecasts', *The British Accounting Review* 50 (1): 16–31.

Brown, S. V. and Tucker, J. W. (2011), 'Large-Sample Evidence on Firms' Year-over-Year MD&A Modifications', *Journal of Accounting Research* 49 (2): 309–346.

Bushee, B. J., Core, J. E., Guay, W., and Hamm, S. J. W. (2010), 'The Role of the Business Press as an Information Intermediary', *Journal of Accounting Research* 48 (1): 1–19.

Caglio, A., Melloni, G., and Perego, P. (2020), 'Informational Content and Assurance of Textual Disclosures: Evidence on Integrated Reporting', *European Accounting Review* 29 (1): 55–83.

Callen, J. L., Khan, M., and Lu, H. (2013), 'Accounting Quality, Stock Price Delay, and Future Stock Returns', *Contemporary Accounting Research* 30 (1): 269–295.

Cazier, R. A. and Pfeiffer, R. J. (2015), 'Why are 10-K Filings So Long?', *Accounting Horizons* 30 (1): 1–21.

Crossley, S. A. and McNamara, D. S. (2011), 'Text coherence and judgments of essay quality: Models of quality and coherence', in Carlson, L., Hoelscher, C., and Shipley, T. F. (eds), *Proceedings of the 29th Annual Conference of the Cognitive Science Society*, Austin, TX: Cognitive Science Society, pp. 1236–1241.

Crossley, S. A., Roscoe, R., and McNamara, D. S. (2011), 'Predicting Human Scores of Essay Quality Using Computational Indices of Linguistic and Textual Features', in Biswas, G., Bull, S., Kay, J., and Mitrovic, A. (eds), *Artificial Intelligence in Education*, Berlin:

Springer, pp. 438–440.

du Toit, E. (2017), 'The readability of integrated reports', *Meditari Accountancy Research* 25 (4): 629–653.

Dyer, T., Lang, M., and Stice-Lawrence, L. (2017), 'The evolution of 10-K textual disclosure: Evidence from Latent Dirichlet Allocation', *Journal of Accounting and Economics* 64 (2): 221–245.

Foltz, P. W., Kintsch, W., and Landauer, T. K. (1998), 'The measurement of textual coherence with latent semantic analysis', *Discourse Processes* 25 (2-3): 285–307.

Grosz, B. and Sidner, C. L. (1986), 'Attention, Intentions, and the Structure of Discourse', *Computational Linguistics* 12 (3): 175–204.

Guay, W. R., Samuels, D., and Taylor, D. J. (2016), 'Guiding Through the Fog: Financial Statement Complexity and Voluntary Disclosure', *Journal of Accounting and Economics* 62 (2-3): 234–269.

Günther, F., Dudschig, C., and Kaup, B. (2015), 'LSAfun—An R package for computations based on Latent Semantic Analysis', *Behavior Research Methods* 47 (4): 930–944.

Higgins, D. and Burstein, J. (2007), 'Sentence similarity measures for essay coherence', *Proceedings of the 7th International Workshop on Computational Semantics (IWCS)*, Tilburg, pp. 1–12.

Hou, K. and Moskowitz, T. J. (2005), 'Market Frictions, Price Delay, and the Cross-Section of Expected Returns', *The Review of Financial Studies* 18 (3): 981–1020.

IIRC (International Integrated Reporting Council) (2021), 'The International <IR> Framework', available at https://integratedreporting.org/wp-content/uploads/2021/01/InternationalIntegratedReportingFramework.pdf.

KPMG (2011), 'Disclosure overload and complexity: Hidden in plain sight', available at http://www.kpmg.com/US/en/IssuesAndInsights/ArticlesPublications/Documents/disclosure-overload-complexity.pdf.

Landauer, T. K. and Dumais, S. T. (1997), 'A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge', *Psychological Review* 104 (2): 211–240.

Lang, M. and Stice-Lawrence, L. (2015), 'Textual analysis and international financial reporting: Large sample evidence', *Journal of Accounting and Economics* 60 (2): 110–135.

Lawrence, A. (2013), 'Individual investors and financial disclosure', *Journal of Accounting and Economics* 56 (1): 130–147.

Lee, Y. J. (2012), 'The Effect of Quarterly Report Readability on Information Efficiency of Stock Prices', *Contemporary Accounting Research* 29 (4): 1137–1170.

Lee, K. W. and Yeo, G. H. H. (2016), 'The association between integrated reporting and firm valuation', *Review of Quantitative Finance and Accounting* 47 (4): 1221–1250.

Li, F. (2008), 'Annual report readability, current earnings, and earnings persistence', *Journal of Accounting and Economics* 45 (2): 221–247.

Loughran, T. and Mcdonald, B. (2011), 'When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks', *The Journal of Finance* 66 (1): 35–65.

Loughran, T. and Mcdonald, B. (2014), 'Measuring Readability in Financial Disclosures', *The Journal of Finance* 69 (4): 1643–1671.

Merkley, K. J. (2013), 'Narrative Disclosure and Earnings Performance: Evidence from R&D Disclosures', *The Accounting Review* 89 (2): 725–757.

Pinker, S. (2014), *The Sense of Style: The Thinking Person's Guide to Writing in the 21st Century*, New York, NY: Penguin.

Rogers, J. L. and Van Buskirk, A. (2009), 'Shareholder litigation and changes in disclosure behavior', *Journal of Accounting and Economics* 47 (1): 136–156.

SEC (Securities and Exchange Commission) (2013), 'Report on Review of Disclosure Requirements in Regulation S-K', available at https://www.sec.gov/files/reg-sk-disclosure-requirements-review.pdf.

Siano, F. and Wysocki, P. D. (2018), 'The primacy of numbers in financial and accounting disclosures: Implications for textual analysis research', available at SSRN: https://ssrn.com/abstract=3223757.

Wolf, F. and Gibson, E. (2005), 'Representing Discourse Coherence: A Corpus-Based Study', *Computational Linguistics* 31 (2): 249–287.

You, H. and Zhang, X. (2009), 'Financial reporting complexity and investor underreaction to 10-K information', *Review of Accounting Studies* 14 (4): 559–586.

Zhou, S., Simnett, R., and Green, W. (2017), 'Does Integrated Reporting Matter to the Capital Market?', *Abacus* 53 (1): 94–132.

Appendix A

Coherence of texts is based on the semantic relatedness between adjoining sentences. The algorithm to compute coherence is as follows:

1) Download annual reports for all South African firms from the Capital IQ database.
2) Convert all annual reports to text format files.
3) Remove all punctuation, numbers, tables, and images from the text file to create the final file for coherence computation.
4) Create a co-occurrence matrix in R using the LSA package in R, using the EN_100K corpus publicly available. The coherence measurement is essentially quantifying the semantic similarity between two sentences. To achieve this, the first step is to create a corpus. A corpus is a set of documents that is used to train the algorithm. Herein, each document from the corpus is read and each sentence in the document is identified as a separate vector of words. Once the vectors are established for all the documents in the corpus, a co-occurrence matrix is created. This matrix consists of the high frequency words in the corpus in each row and the associated sentence words in each column. The distance between vectors of each word identify how far they are from each other, as the overlap of words in the sentences identifies their semantic similarity. The corpus used in this paper is the EN_100K corpus available at http://www.lingexp.uni-tuebingen.de/z2/LSAspaces/. This corpus consists of the entire Wikipedia dump of 2009 and uKWaC, which is a corpus of British text, and is made up of approximately 2 billion words. The semantic matrix available above is created from the 100,000 most frequent (unique) words and the associated words of the sentence are the two words preceding the key word and the three words following the key word.
5) The coherence of the text is now calculated using the 'coherence' function in R. This reads sentences of the supplied text in pairs and calculates the cosine of the vectors that match the key words in the co-occurrence matrix with words in the sentences. This process is described in further detail in Landauer and Dumais (1997).
   a. Local coherence is calculated as the cosine of paired sentences, and global coherence is calculated as the average of the 'local coherence' measures of the document.

Examples of Coherence

The following text has been extracted from the 2013 integrated report of Woolworths Holdings (WHL), a South African multinational retail organisation:

"*The independence and performance of all Non-executive directors is reviewed annually by the Chairman. A formal independence test is performed on* ***those directors retiring by***

rotation at the annual general meeting. *The Memorandum of Incorporation states that Non-executive directors may serve for up to a nine-year period subject to rotation. The Board has the discretion to extend the tenure of* ***a director who has served nine-years*** *after being satisfied that the director is still independent and performing his duties to acceptable standards. Mike Leeming and Chris Nissen would have both served on the Board for nine consecutive years and should retire at the 2013 annual general meeting. The Board, at its discretion, has agreed to extend their tenures for an additional year.* ***These two directors*** *chair the Audit, Risk and Social and ethics committees and it was considered key that their directorships be extended to allow for a smooth transition with the incoming Finance director and management of the Social and ethics committee.* ***Their independence*** *has been assessed and the nominations committee is satisfied that they remain independent.*

The Chairman, Simon Susman, is classified as non-independent by virtue of him having held the position of Group chief executive officer within the previous three years. ***In addition, he holds*** *a number of WHL shares which are material to his wealth. Tom Boardman is the Lead independent director who oversees matters discussed by the Board when the Chairman may, or is perceived to, have a conflict of interest. Board evaluation WHL Board and committee evaluations are performed every two years due to the significant amount of time that is committed to these processes and the feedback/implementation of recommendations. An evaluation was performed in April and May 2013 by an independent service provider. The feedback of the results indicated that the WHL Group's strategic direction is clear, that the Board is competent and that the core Board processes are working well. There is an opportunity to include additional expertise, especially in the fields of information technology and its business application as well as Australian and African retail. The streamlining of board documentation and targeted directors' development programmes are areas which can potentially further enhance the Board's functioning.*"

The complete 2013 integrated report for WHL scores 0.78 on *coherence*, suggesting that the report is 78% coherent on average. This is significantly higher than the mean coherence of approximately 45% in the whole sample. The piece of text shown above is an example of why the coherence is high for this report. In this text, the highlighted parts demonstrate the pronoun density of the text. Pronoun density is the number of third-person pronouns divided by the total number of words. Pronoun density is a proxy for the "givenness" of a text. Givenness captures the proportion of given information to new information in the text. In other words, givenness represents the amount of information that can be recovered from the preceding discourse. Not surprisingly, the linguistics literature documents a strong positive association between the givenness of text and coherence. Consequently, givenness itself has been widely used as a proxy for coherence (Crossley and McNamara, 2011).

To see an example of low coherence, consider the following piece of text, extracted from

the 2013 integrated report of Sentula Mining, a diversified mining company in South Africa:

"*A continuing challenge across the Group is the theft of diesel from vehicles and storage tanks.* ***This increase in the number of litres of diesel used by the Group, not only affects carbon emissions but also the bottom line of the business. Improved internal controls and monitoring continues to be implemented in certain areas in an attempt to curb the problem. Maintenance of vehicles is also a vital component of reducing carbon emissions****. All subsidiaries have strict vehicle maintenance programmes in place, not only to increase the efficiency of these machines but also to maximise their availability.*"

The complete 2013 integrated report of Sentula scores 0.43 on coherence, slightly less than the sample average value. This piece of text highlights potential issues in the coherence of the report. The text is discussing two topics at the same time: theft of diesel and carbon emissions. While the reader could possibly draw a connection between these two, the text itself does not assist the reader to do so. The text begins with identifying a concern related to theft of diesel. However, the next sentence mentions an increased use of diesel by the company. Such a leap of logic potentially lowers the givenness of the text, leading to lower coherence.