Proceedings of the 2nd International Conference on Business and Policy Studies
DOI: 10.54254/2754-1169/7/20230240

Exploration of Earnings Management and Lexical Richness: A Study of Companies in North America

Jingna Liu[1,a,*]

[1] *McGill University, Montreal QC H3A 0G4, Canada*
a. jingna.liu@mail.mcgill.ca
**corresponding author*

Abstract: This paper investigates how the lexical richness of a corporation's annual financial reports connects with earnings management (EM) in all industry sectors (except industries with different accounting methods, such as Utilities and Financial services) and in North America only. In order to explore the above-mentioned relationship, this study introduces new variables – Type-token proportion (TTR). Utilizing the TTR and the square root of TTR (denoted as Unique Index in this paper) to quantify the lexical extravagance of the management discussion and analysis section of the annual report (MD&A), this paper anticipates and finds that firms probably going to have manipulated their earnings to beat the previous year's benchmark and have MD&As with high-level writing proficiency (a kind of complexity). This is coherent with the conclusion of Lo et al.'s paper: good news is inherently easier to communicate, and suggests that words of choice help complicate disclosure and conceal bad news.

Keywords: earnings manipulations, earnings management, lexical richness, readability

1. Introduction

1.1. Introduction

Earnings management (EM) refers to a phenomenon that the management team of a corporation intentionally utilizes accounting policies to produce the firm's fiscal or annual financial statements while concealing some negative facts and figures by manipulating its financial records and choice of words. In other words, a firm's financial statements will present an overly positive view with regard to its financial position or meet the stakeholders' expectations. What's more, it is universally acknowledged that a firm's financial statements are a collection of succinct reports and serve as a wellspring of its financial results/position, cash flows, and the overall status of health at a specific point in time. Consequently, it is an important tool for stakeholders of publicly traded corporations to arrive at conclusions about their equity investments, particularly for those who have voting powers on corporate issues. As a result, many related studies are focusing on 1) various calculation methods of EM, such as mathematical modeling of discretionary accruals using time series data – Jones model (1991); 2) the relationship between EM and readability, such as Gunning Fog Index; 3) EM influence, etc.

However, few scholars mention whether the level of lexical richness of an annual report has any association with EM. Lexical richness, also known as lexical diversity, measures the quality of a

DOI: 10.54254/2754-1169/7/20230240

given text sample and determines the level of writing competency in the most efficient manner [1]. This paper will explore the above-mentioned relationship, "***Will annual reports showing strong writing competency lead to earnings management problem?***", by running regressions on lexical richness ratios and earnings and firm-related variables. This research will focus on the relationship between readability indices and the management discussion and analysis section of the annual report (MD&A), particularly on North American corporations. This paper aims to help the general investors examine the existence of earnings manipulation of a particular firm, which might avoid false investment to some extent.

1.2. Literature Review

At the beginning phase of earnings manipulation identification, the dominating focus of interest was on the corporation's quarterly and annual financial statement data. Beneish [2] developed eight financial statement-related indices – Days' sales in receivables index (DSRI), Gross margin index (GMI), Asset quality index (AQI), Sales growth index (SGI), Depreciation index (DEPI), Sales, general, and administrative expenses index (SGAI), Leverage index (LVGI), Total accruals to total assets (TATA) – and concluded that there without a doubt exists a systematic relationship between the likelihood of EM and variables with respect to financial statements, such as gross margin, receivables, accruals, etc. However, since it only focused on publicly traded companies and manipulators whose earnings were overstated, the study would have restrictions on the detection of privately held corporations as well as those corporations that presented decreasing earnings.

Despite the fact that most researchers were focusing on financial data of a report at that time, Smith and Smith introduced one of the important functions of a firm's financial reports [3], that is, communicating selected financial information to the general investors, which was subsequently measured by readability scores based on Flesch and Dale-Chall formula. They observed that it is future development in readability that enhances the communication function of financial reports in a more general and justifiable manner. Before the 20th century, the sample size of textual-related studies was typically small. For instance, 10 NYSE companies [4], 50 New Zealand firms [5], 2,406 companies [3], etc. In the 20th century, lexical properties of annual financial reports have been paid more attention to and analyzed in depth. Li examined the Management Discussion and Analysis (MD&A) section of annual reports, expanded the sample size to 55,719 firm years [6], and tested the hypothesis based on the Fog index: whether there is any relationship between the readability of the MD&A section, firm performance, and its earnings persistence. The general conclusion is that the more complicated a firm's annual report is, the higher likelihood for firms to conceal less uplifting news. However, the study failed to demonstrate the connection between the intricacy of annual reports and the disguise of horrible news. According to new research, there exists a connection between MD&A sections and earnings management – when firms just beat their past year's benchmark, their annual reports tend to be more complex and somehow more likely to experience earnings management-related problems [7].

Studies conducted by Li [6] and Lo et al. [7] have confirmed that the positive correlation between annual reports' readability and the likelihood of earnings management does exist. However previously mentioned examinations have currently impeccably delineated the relationship, it's hard/expensive to acquire the principal variable – Fog index – as a general investor. Therefore, in this paper, I will use new dependent variables, which are determined through publicly accessible data, and examine if they can supplant the Fog index.

DOI: 10.54254/2754-1169/7/20230240

1.3. Hypothesis Development

This paper presumes that the value of TTR represents the breadth of vocabulary, which can be considered a type of complexity. The closer the TTR ratio is to 1, the greater the lexical richness of the text; the closer the TTR ratio is to 1, the more complicated the texts.

1.3.1 Hypothesis 1

This research assumed that a report with a higher UI will be more complicated for people to decipher, thereby more likely to manipulate their earnings.

1.3.2 Hypothesis 2

This research assumed that a report with a higher TTR will be more complicated for people to decipher, thereby more likely to manipulate their earnings.

2. Data and Methodology

2.1. Research Design

2.1.1 Sample

The data collection process of this paper consists of two-step and is demonstrated below: First of all, this paper utilizes approximately twenty-year financial data from CRSP-COMPUSTAT during the period from 2000 to 2022 and merges it with the Loughran McDonald Master Dictionary dataset. The original sample before merging contains a total of 176,390 observations. For the second (the final) step, this paper has excluded three types of observations. Firstly, the paper removes firms with insufficient or missing data. Secondly, having noticed that firms that belong to industries such as utilities and financial services industries have different operating and financial structures, observations with respect to these industries should be excluded. Last but not least, this paper does not consider firms in a particular industry that has less than 10 fyear-SIC2 pairs. Consequently, the final sample contains a total of 136,130 observations and 5,906 unique firms.

2.1.2 Two Lexical-related Dependent Variables

Type-token ration (TTR) is the total number of unique words (types) divided by the total number of words (tokens) in a given segment of language, which is frequently credited to Mildred Templin, a trailblazer of examination into first language improvement, who gathered jargon information from 480 youngsters matured 3-8 years [1].

$$TTR(N) = V(N)/N \tag{1}$$

N represents the number of tokens, and V represents the number of types.

The portion of unique words in a text represents the variation of words used throughout a text, thus measuring the lexical diversity of a text. Additionally, it has been found to positively correlate with more sophisticated texts [8]. This paper has included both TTR and its square root as dependent variables to examine the lexical density, density as well as sophistication of a given text sample. The formula is shown as follows and the square root of TTR is named as Unique Index (UI) throughout this paper:

DOI: 10.54254/2754-1169/7/20230240

$$\text{Unique Index(UI)} = \sqrt{TTR} = \sqrt{\frac{Number\ of\ Unique\ Words(types)}{Number\ of\ Total\ Words(tokens)}} \quad (2)$$

Method of detecting earnings manipulation. This research sets the firm's last year's earnings as its benchmark instead of comparing it with the industry average. Since only a few cents higher than the previous year can be viewed as a similitude/unchanged, it contains limited information to the general investors. As such, this minuscule change in earnings makes it difficult for the public to accurately decipher, thereby increasing the likelihood of earnings manipulation issues. The study of Lo et al. distinguished firms having a higher probability of manipulating earnings as those corporations with earnings about meeting or merely beating the benchmark. This paper adopted the approach of Lo et al.: they defined a dummy variable called MBE. MBE=1 if a change in earnings per share (ΔEPS) falls in the neighborhood from zero to a small positive number; otherwise, MBE=0.

2.1.3 The Modified Jones Model (MJM) Is Presented Below

$$\frac{TAC_{i,t}}{TA_{i,t-1}} = \beta_0 + \beta_1 * \frac{1}{TA_{i,t-1}} + \beta_2 * \frac{\Delta REV}{TA_{i,t-1}} + \beta_3 \frac{PPE}{TA_{i,t-1}} + \varepsilon_{t-1} \quad (3)$$

However, when calculating the modified Jones model, this paper uses this formula instead:

$$\mathrm{TAC_t} = \beta_0 + \beta_1 * \frac{1}{\mathrm{AT_{t-1}}} + \beta_2 \left(\frac{\mathrm{SALE_t - SALE_{t-1}}}{\mathrm{AT_{t-1}}} - \frac{\mathrm{RECT_t - RECT_{t-1}}}{\mathrm{AT_{t-1}}}\right) + \beta_3 * \frac{\mathrm{PPEGT_t}}{\mathrm{AT_{t-1}}} \quad (4)$$

TAC is total accruals, computed as net profit after tax before extraordinary items less cash flows from operations. $1/AT_{t-1}$ is the inverse of the beginning of year total assets. $PPEGT$ is the total property, plant, and equipment. $RECT$ is total receivables.

For real activities earnings management (RAM), this paper calculated RAM as the negative sum of ($\Delta R\&D\ expense$ plus $\Delta Advertising\ expense$), deflated by beginning total assets, with the help of the approach of Lo et al.[7].

3. Results

3.1. Data Exploration

Table 1: Descriptive statistics summary.

Variable	N	Mean	Std Dev	Lower Quartile	Median	Upper Quartile
Unique_Index	46,127	0.26	0.04	0.24	0.26	0.29
TTR	46,127	0.07	0.02	0.06	0.07	0.08
EPS	46,127	0.63	1.75	-0.36	0.20	1.49
MKT_value	46,127	2475.23	4922.68	64.33	387.44	1948.89
Size	46,127	5.85	2.32	4.16	5.96	7.58
MTB	46,127	2.43	2.12	1.17	1.65	2.69
SpecItems	46,127	-0.01	0.03	-0.01	0.00	0.00
Earnings	46,127	-0.13	0.39	-0.13	0.02	0.07
Loss	46,127	14.53	29.97	0.00	0.00	10.96
benchmark	46,127	-1.08	46.56	-0.13	0.02	0.07
EPS_lag	46,127	0.57	1.66	-0.36	0.19	1.42
epsChg	46,127	-0.08	1.19	-0.52	-0.05	0.33
DA_mJones1	46,127	-0.03	0.13	-0.07	-0.01	0.03
abcfo	46,127	-0.03	0.33	-0.06	0.03	0.10
abpro	46,127	0.00	0.29	-0.14	-0.03	0.09
abdisx	46,127	0.12	0.59	-0.12	0.01	0.21
rem	46,127	-0.09	0.64	-0.35	-0.06	0.19
sic2	46,127	40.07	18.38	28.00	36.00	53.00
DA_mJones1_median	46,127	-0.01	0.01	-0.02	-0.01	-0.01
rem_median	46,127	-0.06	0.01	-0.07	-0.06	-0.06
DA_mJones1_median1	46,127	0.05	0.00	0.05	0.05	0.05
rem_median1	46,127	0.24	0.02	0.22	0.24	0.26

DOI: 10.54254/2754-1169/7/20230240

It shows descriptive statistics summary of our whole sample. For variables, Type-token ratio (TTR) and Unique Index, the mean and median of these two variables are equal, which suggests that variables TTR and UI follow a normal distribution.

Table 2: Correlation matrix table.

Pearson Correlation Coefficients, N = 5906

	Unique_Index	TTR	Size	MTB	Earnings	Loss	epsChg
Unique_Index	1	0.99	-0.49	0.17	-0.20	-0.20	0.04
TTR	0.99	1	-0.48	0.17	-0.20	-0.20	0.04
Size	-0.49	-0.48	1	-0.03	0.46	0.11	-0.03
MTB	0.17	0.17	-0.03	1	-0.52	0.01	-0.01
Earnings	-0.20	-0.20	0.46	-0.52	1	-0.27	-0.14
Loss	-0.20	-0.20	0.11	0.01	-0.27	1	0.27
epsChg	0.04	0.04	-0.03	-0.01	-0.14	0.27	1

It presents the correlation coefficients between dependent and independent variables. Based on the correlation matrix, the higher UI/TTR, the less company size, fewer earnings deflated by total assets, and less loss. Most of the P-values of correlation coefficients among variables are < .0001, thus there is no perfect multicollinearity for regressions. Nonetheless, it is the correlation between Unique Index and TTR (0.9877) that constitutes an exception. Since these two variables are not used in the same models, it will not dirty the regression. *Note.* All coefficients are statistically significant. P-values are smaller than 5%.

3.2. Empirical Findings

Prior to running the regression, this research summarized industries and count the number of firms that have missed the benchmark. Industries "*Chemicals and Allied Products*" and "*Business Services*" are always failed to maintain or beat the benchmark regardless of the 2008 financial crisis and the burst of coronavirus. Additionally, the industry "*Electronic & Other Electrical Equipment & Components*" is a new emerging industry that falls behind previous earnings during the period from 2008 to 2019.

3.2.1 Regression 1

$$Unique\ Index(UI) = \beta_0 + \beta_1 * MBE + \sum \beta_i * Control_i \quad (5)$$

Table 3: Regression results on the dependent variable – Unique Index (UI).

	Independent variable	MBE = 1 when $\Delta EPS \in$		
		[$0, $0.01]	[$0, $0.02]	[$0, $0.03]
		I	II	III
β_1	*MBE*	0.018	0.014	0.013
		(4.18)	(4.22)	(4.13)
β_2	*Earnings*	0.013	0.013	0.013
		(5.54)	(5.47)	(5.47)
	Loss	-0.0002	-0.0002	-0.0002
		(-7.51)	(-7.46)	(-7.51)
	Size	-0.010	-0.010	-0.010
		(-30.46)	(-30.44)	(-30.53)
	MTB	0.004	0.004	0.004
		(9.88)	(9.83)	(9.81)
	SpecItems	0.002	0.003	0.002
		(0.13)	(0.17)	(0.13)
	Constant	0.28	0.28	0.28
	Adj. R-Squared	29.20%	29.21%	29.20%

DOI: 10.54254/2754-1169/7/20230240

From the table above, β_1 is greater than zero, meaning that if this year's earnings fall into the neighborhood of a small positive number, the report will be much richer.

3.2.2 Regression 2

$$TTR = \beta_0 + \beta_1 * MBE + \beta_2 * Size + \sum \beta_i * Control_i \quad (6)$$

Table 4: Regression results on the dependent variable – Type-token proportion (TTR).

	Independent variable	MBE = 1 when $\Delta EPS \in$		
		[$0, $0.01]	[$0, $0.02]	[$0, $0.03]
		I	II	III
β_1	*MBE*	0.011	0.009	0.008
		(4.31)	(4.50)	(4.46)
β_2	*Size*	-0.005	-0.005	-0.005
		(-31.36)	(-31.25)	(-31.39)
	MTB	0.001	0.001	0.001
		(7.57)	(7.50)	(7.50)
	SpecItems	0.046	0.046	0.046
		(6.60)	(6.57)	(6.53)
	Constant	0.08	0.08	0.08
	Adj. R-Squared	26.74%	26.76%	26.74%

From the table above, β_1 is greater than zero, meaning that if this year's earnings fall into the neighborhood of a small positive number, the report will be much richer.

4. Discussion and Conclusion

Our paper explores the relationship between lexical diversity-related ratio and earnings management proxy. This paper has obtained a positive correlation, which is coherent with the findings of Lo et al., particularly when focusing on the small positive neighborhood of earnings management that is bound to have participated in gatherings or genuine exercises the board to increment earnings. All in all, this paper has observed reliable and vigorous proof that corporations that meet or beat the benchmark have more diverse MD&A report, thereby more complex MD&A report.

This paper provides a more accessible method for a general investor to decipher more accurately when the management team of corporations wants to manipulate their earnings.

References

[1] *Li, X., & Zhang, H. (2021). Developmental Features of Lexical Richness in English Writings by Chinese L3 Beginner Learners. Frontiers in Psychology, 12. https://doi.org/10.3389/fpsyg.2021.752950.*

[2] *Beneish, M. D. (1999). The Detection of Earnings Manipulation. Financial Analysts Journal, 55(5), 24–36. http://www.jstor.org/stable/4480190.*

[3] *Smith, J. E., & Smith, N. P. (1971). Readability: A Measure of the Performance of the Communication Function of Financial Reporting. The Accounting Review, 46(3), 552–561. http://www.jstor.org/stable/244524.*

[4] *Lebar, M.A. (1982). A general semantics analysis of selected sections of the 10-k the annual report to shareholders, and the financial press release. The Accounting Review 57, 176–189.*

[5] *Healy, P. (1977). Can you understand the footnotes to financial statements? Accountants Journal, 219–222.*

[6] *Li, F. (2008). Annual report readability, current earnings, and earnings persistence. Journal of Accounting and Economics, 45(2–3), 221–247. https://doi.org/10.1016/j.jacceco.2008.02.003.*

[7] *Lo, K., Ramos, F., & Rogo, R. (2017). Earnings management and annual report readability. Journal of Accounting and Economics, 63(1), 1–25. https://doi.org/10.1016/j.jacceco.2016.09.002.*

DOI: 10.54254/2754-1169/7/20230240

[8] Boven, G. van. (2021, March 22). Measuring readability. Impacter. Retrieved April 17, 2022, from https://impacter.eu/blog/measuring-readability.
[9] Dechow, P. M., Sloan, R. G., & Sweeney, A. P. (1995). Detecting Earnings Management. The Accounting Review, 70(2), 193–225. http://www.jstor.org/stable/248303.

Appendix

Variable Definitions

1. ***ΔEPS*** = change in earnings per share (EPS) from year $t - 1$ to t.
2. ***MBE*** = 1 if ΔEPS falls in the neighborhood from zero to a small positive number; 0 otherwise. (The small positive number is identified in each test.).
3. ***DA*** = discretionary accruals are estimated by the modified Jones model of Dechow et al. [9].
4. ***Earnings*** = Net income at fiscal year-end.
5. ***Loss*** = 1 if *Earnings* < 0.
6. ***Size*** = log(market value of equity) at fiscal year-end.
7. ***MTB*** = $\frac{(market\ value\ of\ equity + book\ value\ of\ liabilities)}{book\ value\ of\ total\ assets}$
8. ***SpecItems*** = amount of special items divided by total assets.