GOTHENBURG UNIVERSITY
SCHOOL OF BUSINESS, ECONOMICS AND LAW

BACHELOR THESIS IN
INDUSTRIAL AND FINANCIAL MANAGEMENT

SPRING SEMESTER 2014

# The Effects of Annual Report Readability on Subsequent Stock Price Volatility
## -An Empirical Study of Swedish Financial Markets

*Authors:*
Marko COTRA 910710
Fredrik JACOBSON 931119

*Supervisor:*
Ted LINDBLOM

June 24, 2014

UNIVERSITY OF GOTHENBURG
SCHOOL OF BUSINESS, ECONOMICS AND LAW

*This page intentionally left blank*

## Abstract

This study investigates the effects of financial reporting on market behaviour. A global trend in the last decade has been the increasing scope of annual reports. This might result in a more complete reporting, but the advantages with increased disclosure should be put in relation to the risk of confusion. Therefore, it is of interest to further examine the effects of increased disclosure. Increased disclosure affects the readability of financial documents, where readability is the ease of which one can understand written text. Understanding how or if the readability of financial disclosures affects market behaviour is both of regulatory interest as well as that of investors.

The aim of this study is to examine how annual report readability affects subsequent stock price volatility in a Swedish context. Using the proxy for readability put forth by Loughran & McDonald (2014), this study tests a hypothesis to determine the relation between the readability proxy and stock price volatility. This is done for annual reports as well as board of directors' reports (förvaltningsberättelse), where the latter is unique to Sweden.

In conclusion, a statistically significant relationship between annual report readability and subsequent stock price volatility is found. However, the economic impact of these findings is limited. A statistically significant relationship between board of directors' report readability and subsequent stock price volatility can not be established.

# Contents

# 1  Introduction

*This chapter first describes a background to the importance of annual report readability. This falls into a problem discussion addressing readability measures and results from previous studies. Finally, the research questions and overall aim of the paper are presented.*

## 1.1  Problem Background

A global phenomenon in the last decade has been the increasing scope of annual reports, especially in the last five years (Lahart, 2014). This implies to a certain extent a more complete financial reporting, but not necessarily a more relevant one. An important aspect is what the trend with increasing scope has resulted in. Do the recipients obtain more knowledge or does more uncertainty arise?

The increasing scope of annual reports has occurred in parallel with investor relations in general becoming a more important subset of corporate communications programs (Hrasky & Smith, 2008). The importance of how information is communicated is a central aspect, since what and when it is presented is less interesting, if the information alone is difficult to take in (Courtis, 2004).

In light of this the European Financial Reporting Advisory Group's (EFRAG) "discussion paper" about "Disclosure Framework" can be understood. The discussion paper aims to decrease the large amount of voluntary disclosure (Deloitte, 2012). Hans Hoogervorst, chairman of the International Accounting Standards Board (IASB), has expressed goals for IASB to decrease the occurrence of unnecessary information (Reuters, 2013). A similar move has been made by the US Securities and Exchange Commission (SEC), which has published "A Plain English Handbook" containing guidelines for making information in financial documents more readable. Consequently, the advantage with increased disclosure should be put in relation to the risk of confusion. Furthermore, the theoretical framework suggests a two-parted relationship between size and benefit.

In accordance with the theory of information asymmetry more voluntary disclosure will result in less information asymmetry being present between managers and the market (Lang & Lundholm, 1993). The effects of this would be lower transaction and agency costs for investors, and would as such be beneficial. However, there are also theories regarding larger annual reports resulting in worse readability. In accordance with the incomplete revelation hypothesis, companies with poor results produce annual reports with lower readability (Bloomfield, 2002). This is also in agreement with the obfuscation hypothesis which states that managers in companies will try to obfuscate bad news by writing longer and less readable texts (Courtis, 2004).

A first approach in assessing this dichotomy then, is to further examine what increased disclosure does result in. For this, textual analysis is suitable in order to determine readability of annual reports.

## 1.2 Problem Discussion

Examining the relationship between financial reports and market behaviour is of interest for both regulators and investors. Firstly, IASB states in their conceptual framework that the purpose of financial reporting is to act as an aid in decision making for current and potential investors and creditors, making annual report readability relevant. With the aim of being used as a basis for decision making, there are requirements on the quality of annual reports. Additionally, studies have found that readability does impact market behaviour making it a concern for investors (Lawrence, 2013; Miller, 2010).

Much research has been performed, trying to determine causal relationships between report readability and the market, looking at among others cost of capital, earnings persistence and stock price volatility (Francis et al., 2008; Li, 2008; Loughran & McDonald, 2014).

Interestingly, previous studies on financial reporting find this relationship to be two-pieced as proposed by theory (Hrasky et al., 2009). On one hand, a positive relationship between scope and stock price volatility has been found (Li, 2008; Loughran & McDonald, 2014). On the other hand, studies investigating the relationship between voluntary disclosure and cost of capital have found a negative relationship, where more information leads to a lower cost of capital (Francis et al., 2008).

The relationship between scope and subsequent stock price volatility also has implications for the efficient market hypothesis. The efficient market hypothesis states that markets react instantly to all new information (Fama, 1970). A significant relationship, however, would indicate that it does not hold true for all markets and under all circumstances.

In order to test the impact of the readability of financial reports on stock prices,it is necessary to find a readability measure which is easy to use and consistent for financial documents. Such a readability measure would make it possible for regulators as well as investors to take readability of financial documents more easily into consideration. Among readability measures, Fog index and Flesch reading ease are the most common (Hrasky et al., 2009). These measures rely on sentence length and number of syllables per word to assess readability.

Loughran & McDonald (2014) present an alternative measure for readability. They use the file size of the financial documents as a proxy for readability. Measuring subsequent stock price volatility after the filing date, they find file size to be a better predictor than both Fog index and other readability measures.

Moreover, previous studies in this field have mainly been conducted in an American context, therefore examining financial reporting under U.S. Generally Accepted Accounting Principles (GAAP) and SEC regulation. By performing this study in Sweden, it is possible to test the readability proxy proposed in Loughran & McDonald (2014) in another context. Additionally, it contributes with knowledge regarding the relationship between annual report readability and market behaviour.

Finally, a significant characteristic of Swedish annual reports is the board of directors' report[1]. The board of directors' report is required by law to be included in the annual report. There are, however, no detailed requirements regarding its length and content. It is therefore of interest to also examine the importance of the board of directors' report; whether it results in increased transparency and understanding of a company's economic standing.

## 1.3   Research Question

It is with these ambiguous results that it becomes interesting to further examine the field. The issue stemming from the problem discussion is two-pieced. First, further examination of the relationship between readability and subsequent volatility is warranted. Secondly, a suitable measure for readability is necessary. Consequently, the research questions are:

- What is the relationship between annual reports' readability and stock price volatility within Stockholm's stock market?

- How is the readability of annual reports adequately measured?

## 1.4   Aim of Study

The main aim of this paper is to examine how annual report readability affects subsequent stock price volatility in a Swedish context. Additionally, this study further aims to investigate suitable readability proxies in a non-U.S. setting, enabling a pragmatic application for investors and regulators.

As mentioned, examining this in a Swedish context makes the board of directors' report of particular interest since it is exclusively a Swedish occurrence. Therefore, relationship between readability and volatility will be tested both on annual reports as a whole, as well as the board of directors' report in isolation.

---

[1]förvaltningsberättelse

# 2  Frame of Reference

*In this chapter different readability measures are presented. Their applicability is discussed, followed by a review of the empirical findings from previous studies. Additionally, an alternative proxy for readability, file size, is presented and evaluated. Finally, a hypothesis is formulated to adress the research question.*

## 2.1  Definition of Textual Analysis and Readability

Textual analysis is a broad topic covering several approaches. Beattie et al. (2004) create a framework for textual analysis by dividing it into three main categories:

- Thematic content
- Readability studies
- Linguistic analysis

The first category looks at what is written and the other two focus on how information is presented. Among these categories, quantitative readability studies are most common in research on financial report readability (Hrasky et al., 2009). However, before discussing how to measure readability a definition of the term is needed.

The meaning of readability tends to differ depending on the context, hence lacking a universal definition. One way of interpreting readability is that of syntactical complexity only. This is in line with Klare's (1963, p. 33-34) definition of readability as "the ease of understanding or comprehension due to the style of writing".

However, there are other definitions, viewing readability in a broader context. From this perspective, not only writing style affects readability, but also target audience and previous knowledge. Following this broader perspective, Loughran & McDonald (2014, p. 11) define readability in a financial disclosure context as "the ability of individual investors and analysts to assimilate valuation-relevant information from a financial disclosure". Having established a definition of readability, a review of the alternative measures is warranted.

## 2.2  Measures of Readability

As previously mentioned, the focal point for readability studies is how information is conveyed, in contrast to the actual content. A more complex text in terms of structure, clauses and sentences will make the information less accessible.

In order to measure readability, different proxies are used to evaluate the complexity. There are numerous variations of readability measures, but their composition is fundamentally the same (Hrasky et al., 2009). The formulae use two variables: sentence length and number of syllables per word. These values are then weighted together. For practical reasons, the results are standardized to fit a preset index in order to allow for easy interpretation. Depending on the specific index, the coefficients and algorithms differ slightly. Below are two of the most commonly used readability measures: Flesch Reading Ease and Fog index (Hrasky et al., 2009). Additionally, Courtis measure of obfuscation is presented, which builds upon the foundation of the other two readability measures.

### 2.2.1 Flesch Reading Ease Index

Flesch Reading Ease is the most commonly used readability formula (Hrasky et al., 2009). The index was created by Rudolf Flesch in 1948 and was derived from a correlation regression on McCall-Crabs standard test lessons in reading and sentence length and number of syllables. The regression coefficients were then standardized to fit a 100-point scale.[2] Ultimately, this gave Flesch Reading Ease equation (Flesch, 1948):

$$206.835 - 1.015 \left( \frac{totalwords}{totalsentences} \right) - 84.6 \left( \frac{totalsyllables}{totalwords} \right)$$

where a higher value denotes higher readability. To give the resulting value context, the values are divided into intervals, which allows for further interpretation. The division varies, with 3-6 intervals common. For example:

| Score | Notes |
| --- | --- |
| 90.0-100.0 | easily understood by an average 11 year-old student |
| 60.0-70.0 | easily understood by 13-15 year-old students |
| 0.0-30.0 | best understood by academics |

### 2.2.2 Fog Index

An alternative measure is the Fog index, developed by Robert Gunning 1952. The derived value from the formula approximates the number of years of formal education an average reader needs in order to comprehend the text. For the formula to be applicable, however, the text needs to be well structured and logically designed (Loughran & McDonald, 2014; Li, 2008). The formula is:

$$Fog = 0.4 \cdot (\text{words per sentence} + \text{percent of complex words})$$

Words per sentence is evidently the average sentence length for the text, while complex words are defined as words with three or more syllables. The retrieved value can then be interpreted as following:

| Score | Notes |
| --- | --- |
| FOG = 18 | Unreadable |
| FOG = 14-18 | Difficult |
| FOG = 12-14 | Ideal |
| FOG = 10-12 | Acceptable |
| FOG = 8-10 | Childish |

---

[2]The highest theoretical value is 120, obtained with two words per sentence and bisyllabic words. The formula has no theoretical lower boundary.

### 2.2.3 Obfuscation

Courtis (1998, 2004) introduce one additional measure of readability by looking at obfuscation. Courtis (2004, p. 291) defines obfuscation as "the simultaneous use of writing with (a) low reading ease and (b) high readability variability"; where readability is calculated using any readability measures. Readability variability is defined as the standard deviation of the readability measure between different passages in the text.

Consequently, obfuscation includes the readability measures in its definition, but is not bound to a specific formula. Therefore, the focal point is not the values from the formula per se, but instead the variation of readability.

## 2.3 Validity of Readability Formulae

The use of these simple formulae is not without problems; the benefits with a pragmatic measure of readability need to be evaluated against its drawbacks. The advantages mainly stem from their simplicity, but also their potential for benchmarking and validation of research questions (Beattie et al., 2004). The drawbacks, however, concern what the measures fail to capture.

### 2.3.1 Supporting Arguments for Readability Formulae

Hrasky et al. (2009) summarize the justification of readability measures in two main arguments. First, they have a long history of use, and they provide a straightforward way to compute readability. An objective measure, with regard to computation, allows for comparison and is a valuable tool for validation of research hypotheses.

Secondly, the measures are mostly used to do relative comparisons, not looking at the absolute value per se. Because the measures are correlated with readability and textual complexity, comparisons using these proxies are still meaningful. Hence, the drawbacks with the simplicity of the measures are mitigated (Courtis, 2004). Finally, Klare (1974-75) defends the use of simple readability formulae of two reasons; firstly, word length is related to speed of recognition. Secondly, sentence length is correlated with memory span. Ceteris paribus, a text with shorter words and sentences is easier to read.

### 2.3.2 Opposing Arguments for Readability Formulae

The criticism for the use of readability formulae mainly concerns their context, simplicity, and what they fail to capture (Hrasky et al., 2009; Beattie et al., 2004). The first drawback with using readability measures is the context in which the indices were developed. Both Flesch Reading Ease and Fog index were developed to measure readability of high school texts for American children in the mid 1900s and have not been recalibrated since (Courtis, 2004). Moreover, the measures were designed for narrative text, not financial disclosures.

Using the formulae in a financial reporting context then is not without issues. Reports tend to get low readability scores because of the nature of the vocabulary used. Typical financial terms are polysyllabic, without having to be difficult for the reader to understand. An example of words identified as complex are: Financial, Company, Interest, Agreement, Including.

Additionally, since the measures are designed for narrative text, some parsing of the text is needed (Li, 2008). Among other, parsing for abbreviations and bullet points is needed in order for the formulae to work properly. Furthermore, Flesch recommends that units of thoughts are measured instead of strict punctuation. Consequently, sub-clauses should be divided into stand-alone sentences. In spite of this, strict punctuation is still most commonly used (Hrasky & Smith, 2008).

Moreover, Hrasky et al. (2009) argue that the readability measures can not be used in neither absolute or relative terms because they are too simplistic. First, they do not consider grammar; text with short words and sentences will give high readability scores regardless the occurrence of illogical word order and lack of verbs. Additionally, since the indices look at sentence length and number of syllables alone, the textual structure, reinforcement of ideas, user-friendliness of fonts, use of supporting imagery and graphs and page layout will not affect readability (Li, 2008; Courtis, 2004).

Finally, there are several aspects which the measures fail to capture. First, they do not consider the reader's background or motivation to reading the text. As a result, the formulae will not differentiate results between target groups, ignoring any possible differences between prior knowledge (Courtis, 2004). In a sample size of 66,707 10-K observations from 1994 to 2011, Loughran & McDonald (2014) illustrate this by presenting the first quartile of the most frequently occurring complex words. The 5 most common words are: Financial, Company, Interest, Agreement and Including.

Furthermore, the use of graphs and charts is not included in the measures. Graphs can explain complex relationships that would otherwise have been difficult to present in writing. Even so, graphs can also be used selectively to obscure information, reinforcing its importance in financial reporting (Hrasky & Smith, 2008).

### 2.3.3  Concluding Remarks

The formulae succeed in measuring complexity, calculated as sentence length and polysyllabic words. Klare (1974-75) finds these bivariate measures a good approximation. Additionally, there is some support from regulators, giving these measures credibility. In U.S. Security of Exchange Commission's Plain English Handbook, both sentence length and number of syllables are identified as important for financial reporting. However, the formulae fail to measure actual readability, both in absolute and relative terms, and should therefore be considered solely as a component in assessing readability (Hrasky et al., 2009).

## 2.4 Empirical Evidence from Readability Studies

There have been numerous studies conducted on financial reporting's impact on the stock market. More specifically, the effect of report readability has been studied. Jones & Shoemaker (1994) summarize previous readability studies and their findings. Looking at studies from 1994 and earlier, the authors conclude that the results are ambiguous without a clear conclusion to be drawn.

Hrasky et al. (2009) perform a similar study, looking at what has been published since 1994. The article reaches the same inference; the results remain ambiguous and contradictory at some points. Furthermore, the article summarizes the methodology and readability measure used in the studies. The result shows that readability formulae are present in all studies, with Flesch Reading Ease, Fog index and obfuscation being the most prevalent ones.

In light of this, Loughran & McDonald (2014) propose file size of the document as an alternative proxy for readability. Given the inconsistent results from past research, a review of their new readability measure is justified.

### 2.4.1 Validation of File Size

Testing both Fog index and file size, Loughran & McDonald (2014) find file size to be a better predictor of post-filing volatility. On a dataset, ranging from 1994-2011 with 66,707 observations, they perform their regression on post-filing stock price volatility as dependant and log (file size) as predictor. After including control variables, the file size coefficient is positive and significant (t-statistic of 4.6). Adding Fog index to the regression, file size remains significant while Fog index is insignificant.

However, the results are not conclusive. Loughran & McDonald (2014) find correlation with the error term, suggesting an important omitted variable. Consequently, there exists some econometric ambiguity regarding collinearity. Nevertheless, their results suggest file size to be a better proxy than alternative measures for readability.

Finally, Loughran & McDonald (2014) examine the economic impact of file size. Looking at the standard deviations of the different variables in the regression model, they find that pre-filing stock price volatility has a larger economic impact than file size. Ultimately, they conclude that file size is a predictor of subsequent stock price volatility, albeit not a primary one.

### 2.4.2 Method of Using File Size

Measuring the file size of a text file is straightforward. Loughran & McDonald (2014) define file size as the byte size of the raw text file of the document. The underlying document used in their study is the 10-K filing required by the SEC. Form 10-K is a comprehensive summary of the company's financial performance with four distinct parts set by SEC.

Using the SEC's EDGAR database, they retrieve the complete submission text file available for all 10-K filings. Therefore, using the text file requires no additional parsing, as the file size is used as proxy.

## 2.5 File Size as a Readability Proxy

The notion of using size as a proxy for readability is not new and has an intuitive explanation. Li (2008) examines the relationship between readability (partially measured as the logarithm of number of words) and firm performance and earnings persistence. The reasoning behind using log (number of words) as a component is that longer reports require more time; hence the processing cost is higher. Loughran & McDonald (2014) also find size to be a better proxy for readability than the above mentioned readability formulae. They argue that obfuscation of information is not likely to occur by the use of long sentences and complex words, but rather by burying the information in longer reports.

Furthermore, this interpretation is well-established in the readability field. It is also in accordance with the incomplete revelation hypothesis proposed by Bloomfield (2002). In the article, Bloomfield tests the hypothesis which states that companies with information to cover produce financial disclosures with lower readability. Li (2008) validates this hypothesis by looking at fog index and number of words as proxies for readability.

However, the relationship between file size and subsequent stock price volatility is not necessarily positive; more disclosure can also lead to less volatility. An alternative use of file size in financial reporting studies is as a proxy for disclosure (Leuz & Schrand, 2009). As such, larger reports are expected to have a negative correlation with subsequent stock price volatility.

This approach has empirical backing as well. Lang & Lundholm (1993) advocate the interpretation that more disclosure lowers the cost of capital and stock price volatility. Furthermore, this has been tested empirically. Botosan (1997) examines the effects of voluntary disclosure on cost of capital and is unable to find a unconditioned statistically significant association. However, for firms with low analyst following, she finds a significant relationship. This is also in line with more recent studies which find a negative relationship between disclosure and cost of capital (Francis et al., 2008).

Finding a significant relationship, regardless of sign, sheds light on how the market assimilates information. Given a relationship between readability and subsequent stock price volatility, this would open up for historical information being used to predict future market behaviour. However, this contradicts the efficient market hypothesis which states that historical information can not be used to predict future stock movements (Fama, 1970).

## 2.6 Hypothesis Formulation

Following from the frame of reference, a hypothesis is formulated to answer the research question. The relationship between readability and stock price volatility will be tested using the following hypothesis (in its nullform):

$H_0$ : *There is no relationship between annual report readability and stock price volatility*

Due to the found dichotomy of file size as proxy for both readability and disclosure, a two-sided hypothesis is chosen. Consequently, the hypothesis does not incorporate what sign the relationship is.

# 3 Methodology

*This chapter presents the methodology used to investigate the research questions, as well as the deductive approach to reach conclusions. It further describes the working process, including sample creation and data gathering. Additionally, deviations from the methodology used in Loughran & McDonald (2014) are discussed. Finally it discusses what actions were made to ensure reliability, replicability and validity.*

## 3.1 Research Philosophy

There are two main approaches to scientific research: deductive reasoning and inductive reasoning (Bryman & Bell, 2013). They differ in terms of the strategy used to reach conclusions. This study used a deductive approach, allowing a hypothesis to be formulated based on the frame of reference. With theories suggesting a two-pieced relationship, it became possible to test which one held true empirically.

Using a deductive approach further impacts the data collection method as the hypothesis determines what data is collected (Bryman & Bell, 2013). Consequently, the data collection method in this study needed to be quantitative.

## 3.2 Working Procedure

The process of this paper consists of four stages: literature review, hypothesis formulation, data gathering and analysis. In order to find a testable hypothesis, a review of the frame of reference was necessary. Next was the procedure for data collection. In this stage, all the necessary data was gathered and processed for the analysis. Finally, in the analysis, the empirical results were discussed and compared to the frame of reference, ultimately leading to a validation of the research question. Additionally, the working procedure was discussed, commenting on possible difficulties.

## 3.3 Literature Review

Initially, previous studies in the field were reviewed, creating a frame of reference. The method used for the literature review was a sequential process presented in Bryman & Bell (2013). The process started with reading known litterature discussing the research question, indentifying key words present. Finally, the key words were used as search words to find additional information from other sources. As proposed by Bryman & Bell (2013), electronic databases were used as they are more reliable.

The e-databases used in this paper were: BSP, Emerald, Science Direct and Elsevier. More specifically, the following search words were used: financial readability, readability annual report, voluntary disclosure, incomplete revelation hypothesis and efficient market hypothesis.

## 3.4   Data collection

The data sources used can be either of primary or secondary nature, where the first refers to data collected primarily by the researcher. Secondary data, on the other hand, refers to data already collected by other researchers or institutions (Bryman & Bell, 2013).

Looking at stock price volatility and readability of annual reports, secondary data was collected for stock market information. To assess readability, annual reports were retrieved. These documents are secondary data, however the information extracted using content analysis provides primary data (Bryman & Bell, 2013).

### 3.4.1   Empirical Data

Using file size in their article, Loughran & McDonald (2014) provide a framework for the parsing needed, variable definitions and regression models. Consequently, after deciding to use file size as readability measure, their analytical models and methodology was adopted.

Examining the relationship in a Swedish context, however, has several implications. The main issue with Loughran & McDonald's (2014) method concerns data gathering. In their study, they use the EDGAR database provided by the SEC. In EDGAR, all firm's 10-K filings' complete submission file is available. Hence, there is no parsing necessary to obtain file size for the report. Below follows a description of the data collection and analysis phase, with comments on deviations from Loughran & McDonald's (2014) methodology.

The first deviation from Loughran & McDonald's study is the use of annual reports instead of 10-K filings. The issue resulting from this is twofold. First, SEC demands 10-K filings to be filed within 60-90 days, depending on firm size. This puts the filing date of 10-Ks well before that of the annual reports in Sweden. Secondly, the content requirements are more restricted in 10-Ks than annual reports, resulting in a more comprehensive overview of the firm (Investopedia).

Nevertheless, without any equivalence of the 10-K filings in Sweden, the annual report is the document where one would find the information contained in the 10-Ks. The year-end report[3] itself would not suffice as substitute for 10-Ks.

Consequently, the procedure for getting file size was altered substantially. Because there is no central database storing all annual reports like the SEC's EDGAR database for 10-K filings, the annual reports had to be retrieved manually: The documents were downloaded in PDF format from Orbis database with missing reports obtained from the companies' web page.

However, because the variable used as readability proxy is file size of the raw text file, additional parsing was necessary to extract the text from the PDF-files. The process included removing file encryption, extracting raw text from the PDFs and printing the byte-size of the raw text file. This was all done using computer aid, giving perfect reliability (Bryman & Bell, 2013). The procedure and software specifications are more thoroughly explained in Appendix C.

---

[3]bokslutskommuniké

Additionally, the board of directors' report had to be located manually within the annual reports, and then it had to be separated, followed by the same procedure for text extraction as annual reports; see Appendix C for further explanation. With the only regulation that board of directors' report is to "...describe in words the most important events that have occurred during the year and after the end of the fiscal year..." (Bolagsverket), it lacks a specific content requirement. Consequently, the board of directors' report's disposition and what is chosen to be included in it differ widely. To refrain from constructing a study-specific content requirement, the individual firm's definition was used when extracting the board of directors' report. Additionally, the firm's definition is also what investors are confronted with. Therefore, using their definition is representative to what investors are presented with.

In addition to the text files, publication dates also had to be determined to calculate post-filing stock price volatility. These dates had to be located manually, looking at press releases through Cision, the companies' web pages and contacting the investor relation departments of the companies.

Ultimately, firm specific data was retrieved from databases, reducing the risk of error in data (Bryman & Bell, 2013). Stock data and control variable values were retrieved from Reuter's software Datastream. Index data for OMXSGI, the index containing all shares listed on OMX Nordic Exchange Stockholm, was retrieved from NASDAQ OMX's website.

### 3.4.2   Sampling Method

Because of the limited accessibility of information, together with the requirement of digital PDF-files for parsing, the time span examined had to be shortened. Additionally, due to the parsing needed, the sample size had to be sufficiently lowered. Consequently, sampling had to be performed. The choices available are probability sampling and nonprobability sampling methods (Bryman & Bell, 2013). In order for generalization, a representative sample of the population is necessary. This is often achieved using probability sampling to avoid bias (Bryman & Bell, 2013).

In spite of this, partly because of accessibility but also due to interest for investors and regulators, a nonprobability sampling method was used. Using judgemental sampling Cortinhas & Black (2012), OMXS60 was chosen as sample. OMXS60 consists of the 60 largest and most traded shares listed on the NASDAQ OMX Stockholm Exchange. Consequently, they are the companies most followed by investors and is of primary interest.

Furthermore, when examining an ongoing and general phenomenom, choice of time span becomes an open question (Bryman & Bell, 2013). Extraordinary events during certain time periods will skew a subset of the sample, for example the financial crisis of 2008-2009. A larger time span has the advantage of diluting these effects. However, as IFRS was adopted in Sweden in 2005, 2005 was selected as start year in order to avoid noise from this implementation in our data.

## 3.5 Analysis

Following the frame of reference and data collection, the first research question was examined using a hypothesis test. In order to test the hypothesis, a significance level was needed (Collis & Hussey, 2009). The chosen significance level determines the probability of a type-I error, which is the incorrect rejection of a true null hypothesis (Cortinhas & Black, 2012). Hence, when deciding significance level, the associated costs and dangers of committing a type-I error need to be evaluated (Collis & Hussey, 2009). According to Collis & Hussey (2009) a 0.05 significance level is generally accepted when conducting research into issues relating to business and management, which also is in line with Bryman & Bell (2013) who state 0.05 to be the highest risk accepted in social science studies.

Therefore, without the possibility for a cost-benefit analysis, a two-step significance level was adopted. Consequently, the hypothesis was tested on a 5 and 1 % significance level.

### 3.5.1 Regression

The hypothesis was tested using a regression on stock price volatility and readability. In accordance with the research question, this was tested on both the annual report and the board of directors' report in isolation. With annual report's readability used to predict subsequent volatility, volatility was used as the dependent variable, with readability as predicting variable. This is best understood from the causality and course of event; the annual reports are published and their content then have an effect on market behaviour. Below follows a definition of the two key components readability and volatility, and then the final regression model is presented.

### 3.5.2 Readability

From the discussion of the validity in readability formulae, and Loughran & McDonald (2014) findings with file size, their proxy for readability was tested in a Swedish context. Hence, to measure the right hand side variable, readability, file size was used as proxy. In order to correct for skewness, the logarithm of file size was used in the regression (Li, 2008).

### 3.5.3 Volatility

In accordance with Loughran & McDonald (2014), stock price volatility was measured as the post-filing root mean squared error (RMSE) [6, 28] days after publication of the annual reports. Six days was chosen to give the stock market approximately a business week to digest the reports, and 28 days to cover a regular business month's trading.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{r}_i - r_i)}$$

where

$$RMSE = \text{the root mean squared error,}$$
$$\hat{r} = \text{the expected stock return for day } i,$$
$$r = \text{the actual stock return for day } i,$$
$$n = \text{number of complete observations.}$$

The expected stock return was calculated by doing a regression on stock return against the market model for [6,28] days. Loughran & McDonald (2014) use CRSP value weighted index as the market model. Without this index applicable in Sweden, a substitute was required. In accordance with capital asset pricing model (CAPM), the Swedish market index OMXSGI was used.

## 3.6   Model Specification

$$RMSE_{jt} = \beta_0 + \beta_1 Log(file\ size)_{jt} + control\ variables_{jt} + \varepsilon_{jt}$$

where

$$RMSE_{jt} = \text{the root mean squared error for firm } j \text{ in year } t,$$
$$Log(file\ size)_{jt} = \text{readability measured as file size for firm } j \text{ in year } t,$$
$$\beta_0, \beta_1 = \text{coefficients retrieved from the regression,}$$
$$\varepsilon_{jt} = \text{error term from the observation for firm } j \text{ in year } t.$$

Using this regression, the two-tailed hypothesis was tested both on the whole annual report document, as well as the board of directors' report isolated. Control variables were added to the regression in order to check the robustness of the relationship. The variables used was adopted from Loughran & McDonald (2014), who choose these variables based on their ability to explain subsequent stock returns. The following control variables were included:

- Firm size - The market capitalization of the firm.
- Book-to-market value - Book value of equity / market capitalization
- Pre-filing RMSE - Historical RMSE, measured by a regression on stock return against the market model.
- Pre-filing Alpha - The alpha retrieved from the regression.

Additionally, both regression models include industry dummies for first level ICB classification and calendar year dummies. For further definition and computation see Appendix A.

### 3.7 Reliability, Replicability and Validity

#### 3.7.1 Reliability

The reliability of a study is dependant on whether the results found were to be the same if the study was to be replicated (Bryman & Bell, 2013). Additionally, for reliability in quantitative studies, it is important to discuss whether the measures are stable or random (Bryman & Bell, 2013).

To improve the reliability in this study, databases were used to retrieve data. In addition to that, all data parsing was done with computer software, securing a stable test. Finally, to test the internal reliability, the data parsing was performed several times.

#### 3.7.2 Replicability

Replicability is closely connected to reliability. To ensure that a study is replicable, the working procedure needs to be thoroughly documented (Bryman & Bell, 2013). Consequently, a replicable study enables other researchers to test whether the study is reliable (Bryman & Bell, 2013).

The working process of this paper was described in detail under section 3.4.1. Performing a quantitative study, the parsing procedure for extracting text files from PDF was also closely documented in Appendix C to enable replicability.

#### 3.7.3 Validity

Validity detemines if the conclusions made from the study can be considered valid and whether they can be generalized (Bryman & Bell, 2013). It is generally divided into four categories:

*Concept validity* tells if the measures used are adequate proxies for the matters being studied (Bryman & Bell, 2013). Additionally, concept validity requires reliability, since a measure which fluctuates randomly can not be of high validity (Bryman & Bell, 2013). To ensure concept validity of this study, measures were adopted from previous studies. After the frame of reference was reviewed, Loughran & McDonald's (2014) methodology was chosen. Their study has gone through a full peer review and has been issued in Journal of Finance, positively impacting validity.

*Internal validity* concerns the causality between the variables being investigated (Bryman & Bell, 2013). Due to the sequential nature of the variables being used, RMSE following the readability, choosing dependant and independant variable was predetermined.

*External validity* is dependant on sample creation, as the external validity determines to what extent the conclusions made can be generalized to a larger context (Bryman & Bell, 2013). Without the possibility for the same scope as Loughran & McDonald (2014), the results found in this paper cannot be of the same magnitude. As such, the external validity of this study was

limited due to sample size. The results are, however, representative for their context and can be indices for the relationship between readability and stock price volatility.

*Ecological validity* requires the method and setting of a study to approximate the real-world that is being examined in order to achieve ecological validity (Bryman & Bell, 2013). Studying secondary data, this study investigates real market behaviour and does .

Finally, for quantitative studies, *measurment validity* is also of interest (Bryman & Bell, 2013). For this study, *face validity* is central as a new measure is developed. To improve face validity in this study, all variables act as substitutes for the ones used in Loughran & McDonald (2014).

# 4 Results

*This chapter presents the empirical results found. First a descriptive overview of file size shows an increasing trend for annual reports, with a more ambiguous result for board of directors' reports. Finally the different regression results are presented with an overview of the most statistically significant variables.*

## 4.1 File Size Trend

A general trend for the chosen time period of 2005-2012 is an increase in the average file size of annual reports as well as board of directors' reports. This is illustrated by Figure 1 and Figure 2. The increase in file size is more consistent with annual reports than with board of directors' reports, where certain years result in a decrease rather than an increase in file size.

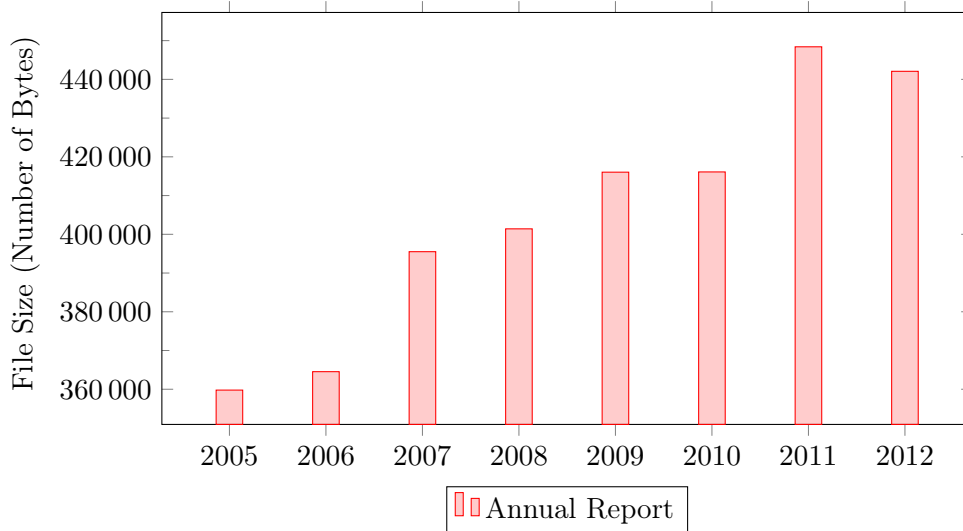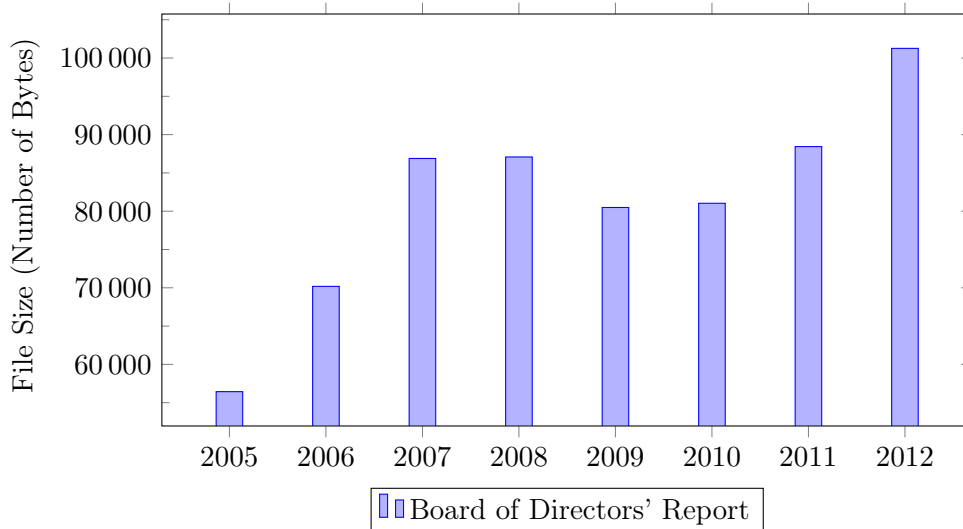Figure 1: File Size Trend for Annual Reports



Figure 2: File Size Trend for Board of Directors' Reports

## 4.2   Regression Results

Below follows the results of the regression analysis in order to determine whether log (file size) is a statistically significant variable in explaining subsequent stock price volatility. Two different regression models are constructed; one using the log (file size) of annual reports as independent variable the other using the log (file size) of board of directors' reports as independent variable. Each regression model includes several control variables (in accordance with Section 3.5.1) in order to check for robustness.

### 4.2.1   Annual Reports

The regression model for determining post filing volatility based on the log (file size) of annual reports is constructed around the data presented in Section 3.5.1. After filtering the data for missing values, the final sample size for annual reports were 362. For further information regarding sample creation see Appendix B. Finally, the resulting regression model is summarized in Table 1.

Table 1: Summary of Regression Model for Annual Reports

| Variable | Coefficient | (t-value) |
|---|---|---|
| Log(File size) | 0.277* | (2.242) |
| Log(Market Capitalization) | -0.163** | (-4.218) |
| Log(Book-to-Market Value) | -0.088 | (-1.407) |
| Pre-filing RMSE | 0.089** | (7.692) |
| Pre-filing Alpha | 0.033 | (1.194) |
| $R^2$ | | 0.543 |
| $F$ | | 19.917 |
| Number of observations | | 362 |

$**/*$ means significance at 0.01 and 0.05 level, respectively

The resulting model has an $R^2$ value of 54.3 %, indicating that the model can account for approximately 54 % of the variation in RMSE present in the sample. The coefficient of log (file size) is significant at a 5 % level (two-sided test), indicated by the corresponding t-value of 2.242. The coefficient has a positive value, indicating that larger financial reports correlate with a larger post filing RMSE. Additionally, the control variables pre filing RMSE and the logarithm of market value capitalization are significant at a 1 % level.

The F-value for the overall model is 19.917, suggesting significant relationships between the response variable and predictor variables. The high t-value associated with pre filing RMSE, signals that the most contributing factor in determining post filing RSME is the historical volatility. As such, the importance of log (file size) of annual reports is comparatively less than the mentioned variable, even though it does contribute to the models.

### 4.2.2 Board of Directors' Reports

The regression model for determining post filing volatility based on the log (file size) of board of directors' reports is constructed around the data presented in Section 3.5.1. After filtering the data for missing values, the final sample size for board of directors' reports were 354. For further information regarding sample creation see Appendix B. Finally, the resulting regression model is summarized in Table 2.

Table 2: Summary of Regression Model for Board of Directors' Reports

| Variable | Coefficient | (t-value) |
|---|---|---|
| Log(File size) | -0.012 | (-0.269) |
| Log(Market Capitalization) | -0.114** | (-3.302) |
| Log(Book-to-Market Value) | -0.107 | (-1.645) |
| Pre-filing RMSE | 0.087** | (7.480) |
| Pre-filing Alpha | 0.029 | (1.054) |
| $R^2$ | | 0.530 |
| $F$ | | 18.357 |
| Number of observations | | 354 |

**/* means significance at 0.01 and 0.05 level, respectively

The resulting model has an $R^2$ value of 53.0 %, indicating that the model can account for approximately 53 % of the variation in RMSE present in the sample. The coefficient for log (file size) is not significant at a 5 % level (two-sided test), indicated by the corresponding t-value of -0.269. The negative sign indicates that larger board of directors' reports correlate with lower subsequent post filing RMSE.

This is directly opposite to the relationship found in Section 4.2.1 with log (file size) of the annual report, which has a positive sign. However, the low t-value indicates that the importance of log (file size) for board of directors' reports is low, providing little additional information to the regression model.

Similar to the regression model in Section 4.2.1, pre filing RMSE and the logarithm of market value capitalization are significant at a 1 % level.

The low significance of log (file size) coefficient for board of directors' reports still results in a model with an $R^2$ and F-value similar (although less than) to the ones achieved in Section 4.2.1 when using the log (file size) of annual reports. This signals that the most contributing factor in determining post filing RSME is historical volatility.

# 5 Analysis

*This chapter initially discusses how the results from Section 4 relate to the relationship between financial report readability and market behaviour. Furthermore, it adresses the use of file size as a readability proxy.*

## 5.1 Analysis of Statistical Results

The hypothesis formulated in Section 2.6 will be used as basis for the analysis of the first research question in Section 1.3. Using the results from Section 4 it becomes possible to address the hypothesis and thus relate it to theories around financial markets and readability.

$H_0$ : *There is no relationship between annual report readability and stock price volatility*

### 5.1.1 Analysis of Annual Reports

For the annual reports, the null hypothesis can be rejected at a level of certainty of 95 %. This indicates that the alternative hypothesis can to a high level of certainty be considered as valid. Annual Report readability by proxy of file size has a positive relationship with (post filing) stock price volatility. This means that the less readable a report is (that is, the larger the file size is) the more volatile the corresponding (post filing) stock price will be.

A significant positive relationship is in line with the results found in Loughran & McDonald (2014). As such, the results found in this study validates the conclusions made in Loughran & McDonald (2014). Together, this can help to reduce the ambiguity regarding readability presented in Hrasky et al. (2009).

The additional factor of the corresponding coefficient being positive is in line with the obfuscation hypothesis. Hence, this provides support to the notion that companies will try to obfuscate bad news by writing longer and less readable texts. In addition to this, the results also provide insight to the incomplete revelation hypothesis by explaining how companies might try to produce reports with lower readability.

Moreover, the positive sign suggests that the relationship between file size and information asymmetry is weak. This is in contrast to the theory prosposed by Lang & Lundholm (1993) where voluntary disclosure per se would reduce information asymmetry.

However, it is important not to overstate the significance of log (file size) in comparison to the other variables present in the regression model. Many of the control variables prove to be highly significant and provide more information to the model than log (file size), mainly the pre filing RMSE and log (Market Value Capitalization) as well as several of the dummy variables accounting for the year the annual reports were filed.

### 5.1.2 Analysis of Board of Directors' Reports

The corresponding null hypothesis for the board of directors' report can not be rejected for the board of directors' report. However, The validity of board of directors' reports as explaining variable is significant when trying to answer the research question proposed in Section 1.3. If the contents of what the board of directors' report includes from year to year differ, then determining whether the size of board of directors' reports correlates with subsequent stock price volatility is inadequate. This is illustrated in Figure 2 (Section 4.1), where the size of the board of directors' reports show no to little consistency. Consequently, the results from the regression model show that log (file size) of board of directors' reports is highly insignificant. However, drawing any meaningful conclusions from regression models based around board of directors' reports should be made with caution, since the validity of the underlying data is questionable.

The issue with the board of directors' report is that it is left open to interpretation beyond the information that is required to be included. This results in several firms changing what is included in the board of directors' report from year to year. Certain annual reports may one year include stock performance as a section under the board of directors' report, and the other year exclude it from the board of directors' report. Consequently, the size of board of directors' reports may vary greatly from year to year simply due to what sections of the annual report are included or omitted from the board of directors' report. As such, this induces a great deal of uncertainty in the regression model for the board of directors' report. This issue can be seen when comparing the general trend of annual report size (bytes of raw text file) during 2005 to 2012 to that of board of directors' report size. The trend of annual reports is a general increase from year to year, with certain years deviating from this trend. Comparatively, the trend of board of directors' reports varies far more from year to year with several years resulting in a significant decrease.

## 5.2 Implications Related to the Efficient Market Hypothesis

Accepting the alternative hypothesis for annual reports as valid has implications about the validity of the Efficient Market Hypothesis. The Efficient Market Hypothesis states that markets react instantly to all available information (Fama, 1970). The alternative hypothesis in conjunction with the regression model for annual reports points to a behaviour where larger annual reports result in more subsequent stock price volatility. A possible explanation to the higher volatility is that it can be seen as a result of the information contained within annual reports being hard to interpret. As such, the subsequent stock price volatility reflects the process of investors interpreting and acting upon the information contained in annual reports.

However, concluding that the Efficient Market Hypothesis is false simply due to higher volatility being linked to annual report size is highly questionable. It is important to consider that the most significant factor in both regression tests was in fact the historical RMSE. The Efficient Market hypothesis states that no future excess returns can be made by using historical information related to stocks.

When dealing with the prospect of generating future excess returns on stocks based on the results obtained in this report it is important to consider the fact that volatility is the key factor which is being studied. Volatility represents the amount and frequency of stock movements. It does not reflect the direction of a stock's movement. This means, under the assumption that the alternative hypothesis is valid, that a subset of the future risk associated with a certain stock can be explained by the size of its annual report. It does not, however, indicate that future excess returns on stocks can be explained by the size of annual reports.

## 5.3 Analysis of Readability Proxy

Analyzing the use of file size as a readability proxy requires a comparison to be made between the regression models in this report to those in the report by Loughran & McDonald (2014). This is due to the fact that the use of annual reports as a basis for readability analysis in this report results in final regression models which are not fully comparable to the ones generated in the report by Loughran & McDonald (2014).

These differences result in different although similar regression models being produced. The similarities between the models produced in this report to those produced in the report by Loughran & McDonald (2014) are that they both aim to investigate the relationship between readability and subsequent stock price volatility. Even though the choice of parameters differs, each chosen parameter serves as a counterpart for the corresponding parameter in the report by Loughran & McDonald (2014).

This means that the underlying principle for both regression models is the same; the readability of reports issued to investors, creditors and regulators has an effect on subsequent stock price. But the difference of the financial document used for this analysis makes it difficult to compare the specific, numerical, results found in this report to those found in the report by Loughran & McDonald (2014).

Even though the specific numerical results cannot be compared, several characteristics can be identified. This is mainly applicable for the regression model produced in Section 4.2.1 for annual reports. Given the high degree of statistical significance for log(file size), as well as the fact that the coefficient is positive, points to similar behaviour to that found in the report by Loughran & McDonald (2014). This further enforces the fact that file size can serve as a good proxy for readability when dealing with financial documents.

An additional factor to consider is that file size was suggested as proxy for readability in a U.S. context. The results found in this report indicate that it can serve as a proxy for readability outside of the U.S, more specifically in Sweden for Swedish annual reports. But an important consideration is the ease of which file size can be used outside of the U.S

Lacking a central database for documents similar to EDGAR results in file size being harder to use. It necessitates the use of more computer aided tools, such as those discussed in Appendix C. It also requires more manual work in terms of fining annual reports and their publication dates. But even though these problems are apparent it is still a comparatively simple proxy to use for

evaluating the readability of financial documents. As discussed in Section 2.1, the conceptual definition of readability is convoluted, covering several aspects related to the use of language, structuring of sentences and so on. With this in perspective, being able to assess readability by looking at file size is still easier to use and automate then to rely on more complex analyses of readability.

# 6  Conclusion

Given the results found and discussed, it is possible to adress the research questions. Answering these questions provides insight to the underlying relationship between the readability of financial reports and the market. The research questions were:

- What is the relationship between annual reports' readability and stock price volatility within Stockholm's stock market?

- How is the readability of annual reports adequately measured?

There is a statistically significant positive relationship between annual report readability and subsequent stock price volatility. This advocates that longer annual reports result in more volatile subsequent stock return. However, this conclusion does not remain valid for board of directors' reports. For board of directors' reports no relationship was found. Finally, file size can be seen as an adequate proxy for readability in annual reports given the statistical significance.

## 6.1  Practical and Theoretical Contributions

These results may serve as indices for the effect of size on subsequent volatility. Additionally, in light of the trend with annual reports growing in size it becomes important to determine what more disclosure results in. This would be beneficial for investors, companies as well as regulators.

The benefit of being aware of this relationship for investors and companies would mainly be risk-related. Companies would need to be more conscious of the length and content of their reports. Meanwhile, investors could use the relationship for input on future volatility.

However, decreasing the size of annual reports is also a regulatory concern. The benefits with more disclosure need to be balanced with the risk of information overload; in some cases less is more.

In general, explaining the characteristics and movements of financial markets is a highly complex task. Assuming that readability by proxy of file size can explain all variations is not reasonable; although the argument made in this report, and in the report by Loughran & McDonald (2014), is that it can provide additional information in addition to that found in more conventional variables and measures. With this in mind, the F-value of 19.917 can be seen as an indication that the model is capable of explaining a subset of the market's behaviour.

## 6.2  Further Research

During the process of creating this report, several interesting aspects were encountered. They were all beyond the scope of this report or discovered during the analysis. These aspects could be interesting to examine and incorporate in further research. Below follows a brief description:

- Define the contents and scope of board of directors' reports more strictly and not adhere solely to individual firms' choice of rubric. This would eliminate the ambiguity regarding what is included in an board of directors' report.

- An increased sample size, in terms of number of firms and time span, would provide more reliability. Additionally, extreme events such as financial crises would not skew the results to such an extent.

- Applying the methodology used in this report, and in the report by Loughran & McDonald (2014), to other financial markets would provide further insight into the relation between the readability of financial documents and corresponding subsequent stock price volatility. Especially examining how country or market specific financial documents adhere to the relations established in this report would also be of interest.

- File size is one way of measuring readability, but several other measures exist. Conducting a more qualitative textual analysis would capture readability in a broader context. Combining both quantitative and qualitative studies would provide useful input for regulators.

# References

Beattie, V., McInnes, B. and Fearnley, S. 2004 "A methodology for analysing and evaluating narratives in annual reports: a comprehensive descriptive profile and metrics for disclosure quality attributes," *Accounting forum*, vol 28, iss 3, pp. 205-236

Bloomfield, Robert J. 2002 ""the Incomplete Revelation Hypothesis" and Financial Reporting," *Accounting Horizons*, vol 16 iss 3

Bolagsverket 2014, Förvaltningsberättelse, Available from: <http://www.bolagsverket.se/ff/foretagsformer/aktiebolag/arsredovisning/delar/forvaltningsberattelse-1.3127>[13 May 2014]

Botosan, Christine A. 1997 "Disclosure Level and the Cost of Equity Capital", *The accounting Review*, vol. 72, iss 3, pp. 323-349

Bryman, Alan and Bell, Emma, 2013, *Företagsekonomiska forskningsmetoder*, Liber, Malmö, Sweden

Collis, Jill and Hussey, Roger, 2009 *Business research : a practical guide for undergraduate & postgraduate students*, 3rd ed, Basingstoke, Palgrave Macmillan.

Cortinhas, Carlos and Black, Ken, 2012 *Statistics for business and economics*, Chichester, Wiley

Courtis, John 1998 "Annual report readability variability: tests of the obfuscation hypothesis", *Accounting Auditing & Accountability Journal*, vol 11, no 4, pp. 459-471

Courtis, John 2004 "Corporate report obfuscation: artefact or phenomenon?", *The British Accounting Review*, vol 36, iss 3, pp. 291-312

Deloitte, *"Joined up writing, Surveying annual reports,"* 2012, available from: <http://www.deloitte.com/assets/Dcom-UnitedKingdom/Local%20Assets/Documents/Services/Audit/uk-audit-joined-up-writing-lowres.pdf> [28 March 2014]

Fama, Eugene 1970 "Efficient Capital Markets: A Review of Theory and Empirical Work". *Journal of Finance*, vol. 25, no. 2, pp. 383–417.

Flesch, R. 1948, "A new readability yardstick", *Journal of Applied Psychology*, vol. 32, no. 3, pp. 221-233.

Francis, Jennifer, Nanda, Dhananjay and Olsson, Per. 2008 "Voluntary Disclosure, Earnings Quality, and Cost of Capital," *Journal of Accounting Research*, vol. 46, iss 1, pp. 53-99

Hrasky, Sue and Smith, Bernadette, 2008 "Concise corporate reporting: communication or symbolism?", *Corporate Communications: An International Journal*, vol. 13, iss 4, pp.418 - 432

Hrasky, S., Mason, C., and Wills, D. 2009 "The Textual Complexity of Annual Report Narratives: A Comparison of High- and Low-Performance Companies", *New Zealand Journal of Applied Business Research*, vol. 7, iss. 2, pp. 31-45

*Investopedia* 2014, 10-K filing. Available from: <http://www.investopedia.com/terms/1/10-k.asp>. [13 May 2014].

Jones, M.J. and Shoemaker, P.A. 1994 "Accounting narratives: a review of empirical studies of content and readability", *Journal of Accounting Literature*, vol. 13, pp. 142-185

Klare, George R., 1963, *The Measurement of Readability*, Iowa State University Press, Ames, Iowa

Klare, George R., 1974-75 "Assessing readability", *Reading Research Quarterly*, vol. X, iss. 1, pp 63-102

Lahart, Justin 2014 "Stop Throwing Book at Investors", *The Wall Street Journal Europe*, 24 March

Lang, Mark and Lundholm, Russell 1993 "Cross-Sectional Determinants of Analyst Ratings of Corporate Disclosures", *Journal of Accounting Research*, vol. 31, iss. 2, pp. 246-271

Lawrence, Alastair, 2013, Individual investors and financial disclosure, *Journal of Accounting & Economics*, vol 56, iss. 1, pp. 130-147

Leuz, Christian and Catherine Schrand 2009 "Disclosure and the cost of capital: Evidence from firms' responses to the Enron shock", *Working paper*, vol. w14897, University of Chicago

Li, Feng 2008 "Annual report readability, current earnings, and earnings persistence," *Journal of Accounting and Economics*, vol. 45, iss. 2-3, p. 221-247

Loughran, Tim and McDonald, Bill, 2014 "Measuring Readability in Financial Disclosures," *Forthcoming: Journal of Finance*

Miller, Brian 2010 "The effects of reporting complexity on small and large investor trading", *Accounting Review*, vol. 85, iss. 6, pp. 2107-2143

*Reuters*, Annual report length, "Time to declutter annual reports, says accounting rule setter," 17 juni 2013, available from: <http://www.reuters.com/article/2013/06/27/accounting-iasb-idUSL5N0F31ST20130627>[28 March 2014]

# Appendix A - Variable definitions

*The following variable definitions were retrieved from Loughran & McDonald (2014)*

| | |
|---|---|
| *Log(file size)* | Natural logarithm of the raw text file in bytes retrieved from parsing the PDFs. |
| *Post-filing RMSE* | RMSE from a market model using [6, 28] days relative to the annual report filing date. Minimum amount of complete observations is 10. |
| *Log(size in $ millions)* | Natural logarithm of the market capitalization of the company one day prior to the publication date. |
| *Pre-filing RMSE* | The RMSE from a market model using [-252, -6]. Minimum amount of complete observations is 60. |
| *Pre-filing alpha* | the alpha calculated using the market model for [-252, -6]. |
| *Log(Book-to-market value)* | Natural logarithm of book-to market value, using market capitalization one day prior to publication date and book value from the most recent year. After trimming for firms with negative book value, the variable is winsorized at the 1% level. |
| *Industry dummy* | Dummy variable for the first level ICB industry classification. |
| *Calendar year dummy* | Dummy variable for the calendar year covered by the report. |

# Appendix B - Sample Creation

The initial OMX60 sample contains several instances of data points which cannot be included in the regression models due to missing data points for corresponding control variables. In addition to this, the amount of dropped data points varies depending on if solely the annual report is being studied or the board of directors' report, resulting in two different final sample sizes.

**Annual Reports Sample Creation**

Table 3 reports the impact of various control variables and data filters on the initial sample when solely investigating annual reports.

Table 3: Impact of Control Variables and Filters for Annual Reports

|  | Dropped | Sample size |
|---|---|---|
| Total number of OMXS60 2005-2012 |  | 480 |
| Available Annual Reports | 9 | 471 |
| Eliminate reports in English/ with sustainability report | 64 | 407 |
| Eliminate reports without known public release date | 31 | 376 |
| Reports whose stock-data is not available for date [6, 28] of post-filing | 1 | 375 |
| Reports whose accompanying year and company does not have known Market Value Capitalization | 8 | 367 |
| Reports whose corresponding stock-data does not have known pre-RMSE | 5 | 362 |
|  |  | 362 |

## Board of Directors' Reports Sample Creation

Table 4 reports the impact of various control variables and data filters on the initial sample when solely investigating the management report.

Table 4: Impact of Control Variables and Filters for Board of Directors' Reports

|  | Dropped | Sample size |
|---|---|---|
| Total number of OMXS60 2005-2012 |  | 480 |
| Available Annual Reports | 9 | 471 |
| Eliminate reports in English/ with sustainability report | 64 | 407 |
| Eliminate reports without known public release date | 31 | 376 |
| Reports whose stock-data is not available for date [6, 28] of post-filing | 1 | 375 |
| Reports whose accompanying year and company does not have known Market Value Capitalization | 8 | 367 |
| Reports whose corresponding stock-data does not have known pre-RMSE | 5 | 362 |
| Reports without distingushiable board of directors' report | 7 | 355 |
| Reports unable to split PDF | 1 | 354 |
|  |  | 354 |

# Appendix C - Parsing of PDF

A necessity in analyzing the size of annual reports as well as the board of directors' reports contained within them is to make all reports comparable to each other. Reports saved in pdf format allow for the ability to include pictures and graphics which inflate the file size and misrepresent the amount of information contained when solely looking at file size (for example due to different resolutions and character encoding). Saving these reports from pdf format to raw text format (UTF-8 format) omits all forms of graphics and only preserves the text present in the reports (this includes text found in tables).

The process of extracting all text contents from annual reports included several steps in order to be performed successfully. The first step was determining whether the pdf file in question was encrypted. Any encrypted files were necessary to be decrypted, making it possible to extract text. The next step was determining which pages in the pdf file contained the board of directors' report and then splitting those contents into a separate pdf file. The final step was locating and extracting all text to a buffer, in order to then transfer the text to a plain text file. Upon creating the plain text file, the file's size on disk could be determined.

The procedure of extracting text from the reports proved to be more problematic than initially thought. Several reports were encrypted, making it difficult to save them as raw text files. The process of breaking encryption as well as saving the reports as raw text resulted in several additional characters being introduced in the reports, such as additional periods or commas. These characters resulted in the file size of the raw text files increasing, but their impact can be seen as miniscule when considering the size of each character (one byte) in comparison to the average file size of the raw text annual reports (approximately 390,000 bytes).

**Decrypting and Extracting Text From Annual Reports**

Several Python scripts as well as MATLAB scripts were written in order to perform most of the tasks listed above. In order to break encryption the site http://www.crackmypdf.com/ was used. The python scripts were written using the pyPDF2 library which allowed for the ability to check whether pdf files were encrypted, splitting pdf files as well as extracting text from pdf files.

One python script was used solely to check which pdf files were encrypted using the pyPDF2 PdfFileReader Class property isEncrypted. Using http://www.crackmypdf.com/ all encrypted files were decrypted. Splitting the board of directors' reports from the annual reports was done by a separate script using the pyPDF2 PdfFileReader Class function getPage(pageNumber) and PdfFileWriter Class function addPage(page) together with a tab-separated text file specifying the pages representing the board of directors' report for each annual report included in the study.

With all necessary files decrypted, board of directors' and all reports separated, a separate python script was used to extract text and save it in a plain text file. This was accomplished with the PageObject Class function extractText() together with the built-in python function write(textstring, encoding), using standard UTF-8 encoding.

With all pdf files saved as raw text in a plain text format a final MATLAB script was used to log the file size (on disk) of all text documents, which was performed with the dir() function. The MATLAB script produced a final log containing the name of each report (consisting of firm name, year and whether it was an board of directors' report or not) as well as the corresponding file size (on disk).