Measuring Readability in Financial Disclosures

Journal of Finance

Tim Loughran

Bill McDonald

University of Notre Dame

UNIVERSITY OF NOTRE DAME
Mendoza College of Business

"Just as the Black-Scholes model is a commonplace when it comes to compliance with the stock option compensation rules, we may soon be looking to the Gunning-Fog and Flesch-Kincaid models to judge the level of compliance with the plain English rules."

- SEC Chairman Christopher Cox in 2007

The comparison is obvious:

Black-Scholes:

$$C(s,t) = N\left(\frac{\ln\left(\frac{S}{K}\right) + \left(r + \frac{\sigma^2}{2}\right)(T-t)}{\sigma\sqrt{T-t}}\right) S \; - N\left(\frac{ln\left(\frac{S}{K}\right) + \left(r - \frac{\sigma^2}{2}\right)(T-t)}{\sigma\sqrt{T-t}}\right) Ke^{-r(T-t)}$$

Gunning-Fog (Fog Index):

$$0.4(w + p)$$

where: w = average # of words per sentence
p = percent of “complex” words (> 2 syllables)

What are we doing?

- We examine the relation between document readability and measures of the information environment (post-filing market model RMSE, earnings surprise, and analyst dispersion) using a large 10-K sample during 1994-2011.

- We define 10-K readability as how effectively management communicates valuation relevant information.

Why is this important?

- Increasing popularity of textual analysis
- Researchers in this area often need a “readability” measure
- SEC continues to emphasize its Plain English initiative
- Most financial/accounting researchers use the Fog Index because it is the most widely recognized measure of readability from other disciplines.

What do we find?

- We show that the Fog Index is an inappropriate measure of readability in the context of financial disclosures.
- We propose using the document file size (i.e., the number of megabytes required to store the document as reported on the SEC website) as a simple, and admittedly imperfect, proxy for readability.
- We show that *file size* relates to post-filing return volatility and other measures of the information environment in a manner consistent with the notion of readability.

What do we find?

- *File size* relates positively and significantly to post-filing date return volatility (market model RMSE), |SUE|, and analyst dispersion after controlling for other variables.

- Fog does not consistently link to these variables and of its two components one is difficult to measure in 10-Ks, while the other is clearly misspecified in applications to financial disclosures.

Literature

- Li (2008)
- Biddle, Hilary, and Verdi (2009)
- Lawrence (2013)
- Miller (2010)
- Dougal, Engelberg, Garcia, and Parsons (2012)
- Lehavy, Li, and Merkley (2011)

The Fog Index

Fog Index = 0.4 (average # of words per sentence +
percent of complex words)

Complex words = words > 2 syllables

Note: Text with Fog Index values greater than 18 is generally considered unreadable.

The Fog Index

Most readability measures, including the Fog Index were derived in the context of grade leveling text books. Consider:

The Cat in the Hat (Dr. Seuss):

> *"The sun did not shine. It was too wet to play. So we sat in the house all that cold, cold, wet day."*

The Jungle (Upton Sinclair):

> *All day long the blazing midsummer sun beat down upon that square mile of abominations: upon tens of thousands of cattle crowded into pens whose wooden floors stank and steamed contagion; upon bare, blistering, cinder-strewn railroad tracks and huge blocks of dingy meat factories, whose labyrinthine passages defied a breath of fresh air to penetrate them; and there are not merely rivers of hot blood and carloads of moist flesh, and rendering-vats and soup cauldrons, glue-factories and fertilizer tanks, that smelt like the craters of hell-there are also tons of garbage festering in the sun, and the greasy laundry of the workers hung out to dry and dining rooms littered with food black with flies, and toilet rooms that are open sewers.*

Ulysses (James Joyce):

> *Molly Bloom's soliloquy – two sentences: (1) 11,281 words (2) 12,931 words.*

Parsing the 10-Ks

- All 10-Ks/10-K405s, excluding amended, from 1994-2011

- Eliminate
  - all ASCII-encoded text (binaries of pictures, pdfs, etc.)
  - all HTML
  - all tables (where >10% of nonblank characters are numbers)

- Parse into words

Parsing for *Fog*

- Two components:

  1. Average sentence length = # words / # sentences
     - # of words from document parse
     - # sentences
       - Remove abbreviations
       - Attempt to identify lists and punctuate
       - Divide number of words by the number of sentence termination characters

Parsing for *Fog*

- Two components:

  2. % complex words (words > 2 syllables)

     - Syllabification
       - "There is no completely agreed-upon definition of a syllable." Jurafsky and Martin (2009, p. 223)
       - Based on 15,000 words manually syllabified, the standard PERL Fathom package is only about 75% accurate.
       - Our algorithm, based on Talburt (1986) and tuning, is > 90% accurate.

What is readability?

- Prior readability research does not provide a precise definition of readability, but acknowledges that its meaning is context dependent.

- Does the SEC intend to make 10-Ks readable for 8th graders?

- Do all those SAT-like words in 10-Ks send you running to the dictionary?

What is readability?

- The definition of Dale and Chall (1948), referred to by Tekfi (1987) as the classic definition and by DuBay (2007) as the most comprehensive definition, specifies “In the broadest sense, readability is the sum total (including interactions) of all the elements with a given piece of printed material that affects the success which a group of readers have with it.”

What is readability?

- Tekfi (1987) concludes that readability is “ensuring that a given piece of writing reaches and affects its audience in the way the author intends.”

What is readability?

- We define readability as the ability of individual investors and analysts to assimilate valuation relevant information from a financial disclosure.

The information environment: Alternative measures

- Post-filing date return volatility
- Absolute value of earnings forecast error (a.k.a. SUE)
- Analyst dispersion

Table I
Sample
Creation

	Dropped	Sample Size
SEC 10-K files 1994 to 2011		188,413
Eliminate duplicates within year/CIK	2,930	185,483
Drop if file date < 180 days from prior filing	585	184,898
Drop if number of words < 2,000	8,298	176,600
CRSP PERMNO match	88,800	87,800
Reported on CRSP as ordinary common equity	4,376	83,424
Price on filing date day minus one ≥ $3	13,338	70,086
Book-to-market COMPUSTAT data available and book value > 0	2,874	67,212
Post-filing date market model RMSE for days [6,28]	346	66,866
At least 60 days data available for market model estimates from event days [-252,-6]	147	66,719
Returns for days 0-1 in event period	12	66,707

Table II
Variable Means
by Time Period
1994-2011

Variable	(1) 1994 to 2002	(2) 2003 to 2011	(3) 1994 to 2011
Readability measures:			
Fog Index	18.44	18.94	18.68
Average words per sentence	22.82	23.27	23.04
Complex words	23.28%	24.09%	23.67%
File size (in megabytes)	0.42	2.51	1.43
Dependent variables:			
Post-filing RMSE	3.45	2.26	2.87
Abs(SUE)	0.27	0.39	0.34
Analyst dispersion	0.14	0.21	0.19
Control variables:			
Pre-filing alpha	0.08	0.05	0.06
Pre-filing RMSE	3.54	2.65	3.11
Abs(filing period abnormal return)	0.04	0.03	0.03
Size (market capitalization) in $ millions	$2,257.56	$3,680.13	$2,946.42
Book-to-market	0.66	0.67	0.66
NASDAQ dummy	0.60	0.58	0.59
# of analysts	4.19	5.60	5.05
Number of observations	34,405	32,302	66,707

Table III
Dependent variable: RMSE
Independent variable: Fog Index
and its components

	(1)	(2)	(3)	(4)
Readability measures:				
Fog Index		0.017		
		(2.04)		
Average words per sentence			0.005	
			(4.02)	
Complex words				-0.006
				(-0.77)
Control variables:				
Pre-filing alpha	-0.913	-0.908	-0.908	-0.912
	(-4.12)	(-4.09)	(-4.10)	(-4.11)
Pre-filing RMSE	0.539	0.539	0.539	0.539
	(12.07)	(12.01)	(12.08)	(12.18)
Abs(filing period abnormal return)	5.057	5.052	5.051	5.056
	(17.52)	(17.57)	(17.57)	(17.53)
Log(size in $ millions)	-0.105	-0.105	-0.105	-0.105
	(-5.45)	(-5.45)	(-5.52)	(-5.50)
Log(book-to-market)	-0.133	-0.133	-0.133	-0.133
	(-2.41)	(-2.41)	(-2.41)	(-2.40)
NASDAQ dummy	0.262	0.262	0.263	0.263
	(3.37)	(3.38)	(3.38)	(3.45)
R^2	46.92%	46.93%	46.93%	46.92%

Table IV First Quartile of Most Frequently Occurring Complex Words in 10-Ks

Word	% of Total Complex Words	Cumulative %	Word	% of Total Complex Words	Cumulative %
FINANCIAL	1.51%	1.51%	ACCOUNTING	0.38%	16.76%
COMPANY	1.44%	2.95%	INCORPORATED	0.37%	17.13%
INTEREST	0.99%	3.94%	INCLUDED	0.37%	17.49%
AGREEMENT	0.78%	4.73%	COMPENSATION	0.36%	17.85%
INCLUDING	0.77%	5.50%	APPLICABLE	0.36%	18.21%
OPERATIONS	0.71%	6.21%	PRIMARILY	0.35%	18.56%
PERIOD	0.71%	6.92%	ACCORDANCE	0.35%	18.91%
RELATED	0.60%	7.52%	SIGNIFICANT	0.34%	19.26%
MANAGEMENT	0.60%	8.12%	SUBSIDIARIES	0.34%	19.60%
CONSOLIDATED	0.58%	8.70%	CUSTOMERS	0.34%	19.94%
INFORMATION	0.58%	9.28%	RESPECTIVELY	0.34%	20.28%
SERVICES	0.55%	9.83%	REGISTRANT	0.34%	20.62%
PROVIDED	0.55%	10.38%	OBLIGATIONS	0.33%	20.95%
PURSUANT	0.55%	10.93%	PROVISIONS	0.33%	21.28%
FOLLOWING	0.54%	11.47%	LIABILITIES	0.32%	21.60%
SECURITIES	0.54%	12.01%	ADDITION	0.32%	21.92%
APPROXIMATELY	0.52%	12.54%	OTHERWISE	0.32%	22.24%
REFERENCE	0.49%	13.03%	PROPERTY	0.32%	22.56%
OPERATING	0.47%	13.50%	EMPLOYEES	0.32%	22.87%
MATERIAL	0.46%	13.96%	BENEFIT	0.32%	23.19%
CAPITAL	0.43%	14.39%	REPORTING	0.32%	23.51%
EXPENSES	0.42%	14.81%	PRINCIPAL	0.31%	23.82%
CORPORATION	0.40%	15.21%	DEVELOPMENT	0.31%	24.13%
OUTSTANDING	0.40%	15.61%	REVENUE	0.30%	24.43%
ADDITIONAL	0.39%	16.00%	EQUITY	0.30%	24.73%
EFFECTVE	0.38%	16.38%	INSURANCE	0.30%	25.04%

Table V
Correlations of Alternative Readability Measures

	Log(file size)	*Fog Index*	*Average words per sentence*	*Complex words*	*Common words*	*Financial terminology*	*Vocabulary*
Fog Index	0.367						
Average words per sentence	0.316	0.885					
Complex words	-0.015	-0.089	-0.542				
Common words	-0.619	-0.465	-0.572	0.385			
Financial terminology	-0.407	-0.301	-0.372	0.254	0.781		
Vocabulary	0.668	0.497	0.596	-0.377	-0.970	-0.724	
Log(# of words)	0.712	0.560	0.652	-0.384	-0.916	-0.615	0.946

Table VI
Dependent variable: RMSE

A Comparison of Log(file size), Fog Index, and the Components of Fog Index

	(1)	(2)	(3)
Readability measures:			
Log(file size)	0.073	0.069	0.076
	(4.60)	(4.25)	(3.36)
Fog Index		0.006	
		(0.73)	
Average words per sentence			0.003
			(0.75)
Complex words			0.010
			(0.67)
R^2	46.96%	46.97%	46.97%

Table VII

Robustness

Dependent variables:	(1) *Abs(SUE)*	(2) *Abs(SUE)*	(3) *Abs(SUE)*	(4) *Analyst Dispersion*	(5) *Analyst Dispersion*	(6) *Analyst Dispersion*
Readability measures:						
Log(file size)		0.046			0.023	
		(5.53)			(3.51)	
Fog Index			-0.003			-0.000
			(-0.82)			(-0.02)
Control variables:						
Pre-filing alpha	-0.365	-0.361	-0.366	-0.270	-0.268	-0.270
	(-4.68)	(-4.68)	(-4.69)	(-4.99)	(-4.99)	(-4.99)
Pre-filing RMSE	0.117	0.115	0.117	0.088	0.087	0.088
	(5.47)	(5.36)	(5.48)	(4.84)	(4.80)	(4.84)
Abs(filing period abnormal return)	0.960	0.958	0.962	0.514	0.510	0.514
	(4.84)	(4.84)	(4.85)	(4.36)	(4.33)	(4.33)
Log(size in $ millions)	-0.054	-0.061	-0.054	-0.022	-0.026	-0.022
	(-6.21)	(-6.71)	(-6.22)	(-4.15)	(-4.78)	(-4.16)
Log(book-to-market)	0.104	0.099	0.104	0.065	0.062	0.065
	(4.85)	(4.64)	(4.85)	(4.09)	(3.96)	(4.10)
NASDAQ dummy	-0.089	-0.086	-0.089	-0.053	-0.051	-0.053
	(-6.56)	(-6.58)	(-6.53)	(-3.81)	(-3.87)	(-3.81)
# of analysts	-0.002	-0.002	-0.002	0.006	0.006	0.006
	(-1.11)	(-1.37)	(-1.11)	(3.35)	(3.34)	(3.35)
Sample size	28,434	28,434	28,434	17,960	17,960	17,960
R^2	23.30%	23.54%	23.30%	25.34%	25.52%	25.34%

Table VIII

Complexity

Dependent variables:	(1) *Post-Filing RMSE*	(2) *Abs(SUE)*	(3) *Analyst Dispersion*
Log(file size)	0.084	0.044	0.022
	(4.84)	(4.93)	(3.49)
Business segment index	0.156	-0.045	-0.009
	(3.80)	(-2.14)	(-0.87)
Observations	50,739	22,783	14,516
R^2	47.27%	21.83%	23.98%

Table IX

Alternative Readability Measures – 24 separate regressions

	Dependent Variable		
	(1)	(2)	(3)
Readability Measures	*Post-Filing RMSE*	*Abs(SUE)*	*Analyst Dispersion*
Log(file size)	0.073	0.046	0.023
	(4.60)	(5.53)	(3.51)
Fog Index	0.017	-0.003	-0.000
	(2.04)	(-0.82)	(-0.02)
Average words per sentence	0.005	0.002	0.002
	(4.02)	(2.23)	(2.34)
Complex words	-0.006	-0.014	-0.009
	(-0.77)	(-5.75)	(-4.08)
Common words	-1.295	-0.614	-0.437
	(-4.56)	(-5.49)	(-4.47)
Financial terminology	-8.601	-1.460	-0.906
	(-4.34)	(-2.68)	(-2.51)
Vocabulary	7.826	4.094	2.835
	(4.72)	(6.31)	(5.68)
Log(# of words)	0.086	0.062	0.041
	(4.27)	(6.55)	(4.79)
Number of observations	66,707	28,434	17,960

Conclusion

- The *Fog Index* doesn't work for financial disclosures.
- We show why.
- We propose an alternative measure for readability – Log(file size).
- *File Size* is imperfect, but it is easy to calculate and relates to all of the information variables we test in a way consistent with the notion of readability.

Conclusion

- As a simple proxy for readability use file size. If you are focusing on other specific attributes of readability consider some of the alternative measures we examine.

Conclusion

- SEC – The SEC should focus less on style—which is undifferentiated in 10-Ks—and instead encourage managers to write more concisely.