Measuring Qualitative Information in Capital Markets Research

Elaine Henry
ehenry@exchange.sba.miami.edu
School of Business
University of Miami

Andrew J. Leone
a.leone@miami.edu
School of Business
University of Miami

November 2010

Abstract

A growing stream of research in accounting and finance tests the extent to which the tone of financial disclosure narrative, also referred to as its qualitative information, affects security prices, over and above the disclosed financial performance. These studies typically measure tone by counting the relative frequency of positive versus negative words in a given disclosure such as earnings press releases. Critical to word-frequency based analysis is the list of words deemed to be positive or negative. Because general wordlists (GI or Diction) likely omit words that would be considered positive or negative in the context of financial disclosure and include words that would not, we expect that these general wordlists be less powerful for hypothesis testing compared to wordlists specifically for the domain of financial disclosure (FD). Using a sample of 29,712 earnings press releases, we find that the context-specific FD wordlist produces a more powerful predictor of market reaction than the general wordlists. Additionally, in smaller samples – demonstrated here with 250 regressions using randomly-selected subsamples ranging in size from 50 to 2,000 – the domain-specific FD wordlist retains predictive ability, with rejection rates exceeding 97 percent for samples of 2,000 while the rejection rates for the general wordlists are less than 30 percent. The FD wordlist also performs better than an alternative, domain-specific wordlist. Overall, our findings indicate that the domain-specific FD wordlist provides an alternative, more powerful measure of tone for capital markets researchers. Finally, we show that equal weighting of word occurrences is more intuitive, easier to implement, and more amenable to replication than alternative sample-dependent weighting methodologies advocated by certain concurrent research.

We thank workshop participants at the University of Colorado, Bill Mayew, DJ Nanda, Sundareash Ramath, Steve Rock, and Peter Wysocki for helpful comments and suggestions. We also thank the University of Miami School of Business for financial support.

*"**Beat***
–verb
1. to strike repeatedly…
4. to overcome…to surpass"
Merriam-Webster online dictionary

1.0 Introduction

A large body of research has examined the relation between *quantitative* financial disclosure and security prices, but a growing stream of capital markets research now examines the impact of *qualitative* financial disclosure, such as tone, on security prices.[1] Although research on qualitative disclosure dates back to at least the early eighties (e.g., Frazier et al. 1984), advances in computing power and software have made the application of computational linguistics accessible to researchers on a large scale. It is now possible for accounting researchers to study qualitative disclosure along many of the same lines that quantitative disclosure has been studied.[2]

A popular area of qualitative research in accounting and finance has been to measure the tone of various disclosures and assess its impact on the capital markets. For the most part, the tools and techniques accounting researchers use for qualitative research have been adopted from linguistics and related fields. Though there are clear advantages to adopting methodologies from other fields, where they have been widely tested and well refined, it is also important to consider whether such methods should be altered or refined for the particular setting being studied. In this study, we evaluate measures of tone used in prior accounting and finance research and recommend an approach that is intuitive, transparent, simple to implement, and the most powerful in the context of financial disclosure.

[1] See Kothari (2001) for a review of the literature on quantitative financial disclosure.
[2] For example, researchers could examine how certain incentives induce managers to bias qualitative disclosure; how timely qualitative information is impounded into security prices; the extent of conservatism in qualitative disclosure; and sources of cross-sectional variation in the market reaction to qualitative disclosure.

Researchers typically measure the tone of disclosures based on frequency counts of certain words that appear in the disclosure.[3] The two choices researchers make that most influence the "tone score," are the wordlists and the weighting schemes. The wordlists are those words deemed to reflect "positive" and "negative" tone. The weighting schemes determine the weights assigned to each of the positive and negative words identified in a particular disclosure. The most common weighting scheme is equal weighting, where each occurrence of a word in a word list is assigned the same weight. For example, a 2,000 word document containing 125 "positive" words and 75 "negative" words would receive a tone score of (125-75)/(125+75)= 0.4. Both alternative wordlists and alternative weighting schemes proposed in the literature are evaluated in this study.

Three commonly used wordlists in capital markets research to measure whether the tone of financial disclosure are: 1) a wordlist developed in Henry (2006, 2008)[4] specifically for use in the domain of financial disclosure; 2) a wordlist from Diction software developed and used by Roderick Hart, a specialist in politics and mass media; and 3) a wordlist from the General Inquirer (GI) program developed and used by Philip Stone, a specialist in social psychology.[5] The wordlists will be referred to as the as the FD wordlist, the Diction wordlist, and the GI wordlist, respectively. In accounting research, Henry (2006, 2008) uses the FD wordlist to examine earnings press releases, Davis et al. (2007) and Demers and Vega (2008) use the Diction wordlist to examine earnings press releases, and Kothari et al. (2009) use the GI wordlist to examine a variety of corporate disclosures.[6] All of these studies employ equal weighting. We examine these particular wordlists because of their relatively wide use in existing accounting and finance

[3] Earlier researchers (Francis et al. 2002; Hoskin et al. 1986) subjectively judge the tone of each disclosure rather than employing computational linguistics to measure tone. An alternative approach to measuring tone with computational linguistics, discussed in a subsequent section, relies on learning algorithms that first develop rules based on a manually-coded, training data set and then apply those rules to measure the qualitative information in a larger data set (e.g. Li 2010, Antweiler and Frank 2004).

[4] The choice of wordlist is independent of the choice of text-processing software. For example, a study can use the Diction software's built-in wordlists but process the texts using other commercial software or using Perl programming. Henry (2006, 2008) uses customized wordlists, available in both publications, and employs Diction software to process texts, in contrast with other research using the wordlists "built in" to the Diction software.

[5] The wordlist used in the General Inquirer text-processing program, the Harvard IV-d dictionary, is available at: http://www.wjh.harvard.edu/~inquirer/spreadsheet_guide.htm.

[6] The methodology of Henry (2006, 2008) has also been used in Gordon et al. (2008) and Sadique et al. (2008). The methodology used in Kothari et al. (2009) has also been used in Tetlock (2007), Tetlock et al. (2008) and Engelberg (2007).

research. We also examine a fourth wordlist developed in Loughran and McDonald's (2011) analysis of 10-K filings, a study which used both equal weighting and an alternative weighting scheme. We will refer to wordlist from this study as the LM wordlist.

Disclosures in earnings announcements are the focus of this study for several reasons. First, unlike periodic SEC filings such as 10-Ks and 10-Qs, earnings press releases are voluntary disclosures. Although companies must furnish any press release announcing earnings to the SEC on a Form 8-K, there is no requirement that a company issue an earnings announcement.[7] Furthermore, 10-K filings contain numerous disclosures and attachments that are boilerplate and sometimes prepared entirely by a third party (e.g., audit opinions, loan agreements, etc.). Second, research consistently demonstrates market reaction on earnings announcements dates, but not consistently on 10-K filing dates.[8] By using earnings announcement dates, we narrow our focus to a critical period of market reaction to a key type of voluntary disclosure. Third, unlike financial narrative by financial journalists or analysts, earnings press releases are disclosures authored by the company.[9] Thus our study offers results that are more relevant for a researcher interested in the impact of disclosure (as opposed to media or analyst coverage).

We predict that the domain-specific FD score will outperform GI and Diction scores as a measure of tone for the following reasons. First, numerous words have meanings in the context of financial disclosure that differ from the meanings assumed by GI and Diction. For example, GI classifies "shares," and "outstanding," as positive words, and "tax" as a negative word.[10] Second, certain words commonly considered positive or negative in a financial disclosure setting, such as "record," are excluded from the GI and Diction wordlists.

---

[7] As stated in SEC Regulation G, the amendment to Form 8-K, Item 12, "does not require that companies issue earnings releases or similar announcements. However, such releases and announcements will trigger the requirements [to furnish the press release to the SEC]" (SEC 2003).

[8] See Easton and Zmijewsk (1993) for a discussion of evidence of little abnormal stock price change around 10-K filing dates. Even in the post-Edgar era, significant market reaction surrounding SEC filings occurs only in limited circumstances [market reaction to 10-Q/QSB/KSB reports occurs only when such reports announce earnings for the first time, i.e., when there has been no previous earnings announcement, and significant market reaction surrounding 10-K filing dates occurs only at certain calendar periods] (Xeujun Li and Ramesh 2009).

[9] Studies that use computational linguistics to study disclosures in financial press include Tetlock et al. (2008) and Kothari et al. (2009).

[10] Use of a domain-specific wordlist mitigates polysemy (the capacity for a single word to have multiple meanings), one of the key generic issues in computational linguistics.

To evaluate the alternative tone measures, we collect a sample of 29,712 earnings press releases issued between 2004 and 2006 from the SEC Edgar site. For each press release, we compute tone measures using each of the alternative wordlists. We then test the market reaction to the earnings news and to the tone of the release, alternately using FD, Diction, GI, and LM tone scores in a short-window event study incorporating each of the alternative the tone scores (i.e., a cross-sectional regression of announcement date abnormal returns on unexpected earnings and tone). Consistent with our expectations, Vuong (1989) tests confirm that empirical models incorporating the FD score better explain market reaction to earnings announcements than models using the GI and Diction tone scores. Further, the economic significance of the FD score is more than twice that of either the Diction score or the GI score. Specifically, a one standard deviation change in the FD score corresponds to a .08 standard deviation in *CAR* versus 0.03 for Diction and GI scores. We also find that empirical models incorporating the FD score better explain market reaction to earnings announcements than models using the LM tone score when the model includes an indicator variable for loss-making firms.

We also test the relative predictive power of the four scores decomposed into positivity (based on a count of all positive words in the disclosure) and negativity (based on a count of all negative words), and using alternative specifications (e.g., scaling and log transformations). The decomposed positive and negative scores of all tone measures are generally statistically significant, with the expected signs, in regressions of returns on unexpected earnings and tone scores. The decomposed FD, GI and LM scores, unlike the decomposed Diction score, are incrementally informative when both are included in regressions of returns on unexpected earnings and tone score. Similar to Tetlock et al. (2008) we find a smaller market reaction to positive tone compared to negative. Specifically, although the coefficients on both the negative and positive FD scores are significant, the economic significance of negativity more than 20 percent higher than that of positivity. This is consistent with investors placing less weight on positive qualitative disclosure than on negative (Hoskin et al. 1986; Hutton et al. 2003; Tetlock 2007; Tetlock et al. 2008).[11]

[11] Hoskin et al. (2002) find that the market responds to both positive and negative prospective officer comments, but the statistical significance is greater for negative prospective comments than for positive. Hutton et al. (2003) find that

In addition to our primary tests, we test the relative power of the tone scores in smaller samples. We create randomly-selected subsamples (with replacement) ranging in size from 50 to 2,000. The null is rejected 98% of the time in random samples of 2,000 observations. In contrast, the null is rejected only 27% and 31% of the time using the non-domain specific Diction and GI tone scores, respectively. We conduct similar tests for decomposed negative and positive scores and find consistent conclusions overall. These results indicate that in contexts where researchers are limited to smaller samples, Type II errors can be avoided by employing the domain-specific FD tone score. Examples of specific contexts in which capital markets researchers focus on smaller samples include subsamples of firms that exceed analysts' forecasts (identified in Brown and Caylor 2005 as perhaps the most important earnings benchmark) or a subsample of high-litigation risk firms that have been the focus of previous capital markets research (e.g. as used in Ajinkya et al. 2005 and Francis et al. 1994).

In addition to our large sample tests, we conduct a detailed analysis of the group of press releases where the tone scores diverge the most. This analysis helps us to understand why the FD score dominates the other measures. As noted above, GI and Diction contain a number of words that are not suitable for the specific domain of financial disclosure and also ignore certain words commonly used to imply positive results. The word "record," which is excluded from the GI and Diction lists, is included in the FD positive wordlist because it is a typical adjective in earnings press releases, referring to "record profits" or "record revenues."[12]

Finally, we advocate the use of equal weighting schemes based on word frequencies (*wf*) and caution against the use of sample-dependent document-weighting methods, such as the one advocated in Loughren and McDonald (2011). The method advocated by Loughren and McDonald combines a sample-dependent, inverse document frequency weighting (*idf*) with word frequency counts to create the word-frequency-inverse document frequency measure (*wf-idf*). The *wf-idf* measure is commonly used in

---

the market consistently responds to bad news qualitative disclosures but only to good news qualitative disclosures when accompanied by verifiable forward-looking statements.

[12] The FD wordlist specifically excludes the word "recorded" because that form of the word signifies the past tense of the verb "record."

information retrieval algorithms (e.g., Google or other search engines) where the objective is essentially to rank a group documents according to their relevance to a specified search topic, but the method does not logically transfer to the measurement of tone.[13] The inverse frequency method, and others like it, down-weight words that are common in the sample of documents, and up-weight less common words. As an example, suppose a user enters the search terms "the" and "golfer." Because the word "the" appears in virtually all documents being searched, the word is down-weighted to the point where it is essentially ignored and the word "golfer" gets almost all of the weight.[14]

In measures of qualitative information in financial disclosure, the magnitude of an equally-weighted tone measure is increasing in the proportion of all words appearing in a document that are classified as positive versus negative. Thus, the probability that an equally weighted measure (a *wf* measure) correctly approximates the tone of a particular disclosure is primarily a function of whether words have been correctly classified as signifying positive and negative information. In measures that add a sample-dependent *idf* weighting (*wf-idf* measures), the magnitude of the measure is decreasing in the frequency with which any given positive or negative word appears in any document within the specific sample of documents, whether or not the word has been correctly classified. As a consequence, an *idf* weighting negates the impact of misclassified words that appear in many documents in the sample (e.g. "tax" in the GI wordlist); however, an *idf* weighting also diminishes the impact of frequently-occurring but correctly-classified words and increases the impact of infrequently-occurring misclassified words. The *idf* methodology is highly effective, and thus widely used, in the domain of "Search" because it helps rank a group of documents according to their relevance to a search term; however, it will not necessarily serve to increase the probability that a measure correctly captures the content of a particular document. Instead the methodology captures the distinctiveness of the negative or positive content relative to the overall sample.

[13] This method is more commonly known as "term-frequency-inverse document frequency (tf-idf), but for consistency with our use of "word" versus "term" in this paper, we replace the word "term" with "word."

[14] Most information retrieval tasks actually omit words that are extremely commonly occurring because they add nothing to measuring content or to discriminating between documents' relevance to the search. Words such as "the" can generally be ignored in keyword-oriented information retrieval without having much impact on retrieval accuracy; such words are known as "stop words" (Manning and Schutze, 2002).

While we do find that using an *idf* weighting provides some improvement to the power of tone measures for some of the randomly-selected samples, the improvements are not uniformly consistent and the amount of the improvement is modest compared to the overall negative aspects of using *idf* weightings. The most negative aspect of using *idf* weightings in measuring the tone of disclosure for capital markets research is that the resulting tone measures are completely determined by the other documents contained in the sample of documents. For any given document, its equally weighted word-frequency tone measure is document-specific and is unrelated to the composition of the sample, but its *wf-idf* tone measure is completely dependent on the composition of documents in the sample and can vary dramatically depending on which and how many other documents are in the sample. In other words, adding an *idf* weighting to a frequency-based measure virtually guarantees that a tone measure would fail one of the fundamental tests of measurement validity. Consequently, we advocate the use of (a) domain-specific wordlists (e.g., FD wordlist) and (b) equal weighting, both of which lead to measures that are more intuitive, more powerful, easier to implement, and – importantly – more amenable to replication.

The remainder of this paper is organized as follows. Section 2 describes alternative measures of qualitative information in capital markets research and explains our motivation for examining these alternatives. Section 3 describes sample selection and defines the specific tone measures we will examine. Section 4 presents descriptive statistics and the main results of our empirical tests of the relation between market reaction to earnings announcements and the alternative measures of the tone of the announcement's narrative. Section 5 reports results of empirical tests of the relative power of the alternative tone measures. To identify reasons for the differences in scores, Section 6 presents an analysis of the group of press releases where the tone scores diverge the most. Section 7 concludes.

2.0 Measures of Qualitative Information in Capital Markets Research

Early capital markets research assessing the market implications of qualitative financial disclosure relied on human coders making item-by-item subjective assessments of tone (Francis et al. 1994; Hoskin et

al. 1986; Lang and Lundholm 2000; Francis et al. 2002). As Core (2001, 452) wrote in a review of capital markets research:

> I conjecture that researchers can substantially lower the cost of computing these [disclosure] metrics by importing techniques in natural language processing from fields like computer science, linguistics, and artificial intelligence. . . . Natural language processing programs could be also used to create proxies for the 'tone' of disclosure and proxies for the precision and bias of the information that is conveyed.

Today, applied computational linguistics (comprising various tasks including content analysis and information retrieval) is a well-established literature in many disciplines. A widely-used reference book on content analysis notes, "it's rare to find a text content analysis today that does *not* use computer analysis" (Neuendorf 2002).[15] One of the most basic tools in computational linguistics captures the content of a document with a count of the relative frequency with which particular words appear.[16] This approach treats a document as a "bag of words," a list of the words occurring in the document and the number of times that word appears (i.e., a vector of word-frequency counts). Word frequency counts are used not only to characterize the content of textual communication, but also to measure the tone and numerous other attributes. To measure tone, the approach generally involves first constructing wordlists of positive and negative words and then counting the number of times these words appear in the text being analyzed. The counts of positive and negative words can be scaled by total words, or can be combined in a single measure. As we define it here, the tone Score is (Positive–Negative)/(Positive+Negative). Now that the tone is quantified, researchers can use the scores to test hypotheses about how tone relates to other observable outputs and ultimately about how tone impacts decision-making.

Among the word lists that have been used in capital markets research, two widely-used wordlists are from other domains: Diction and GI. Diction's wordlists were developed in the domain of political communications and are included in the commercially available software Diction 5.0. (Note that the Diction

[15] The term content analysis, defined as ""a research technique for the objective, systematic, and quantitative description of manifest content of communications" (Berelson, 1952, 74) is an extremely broad term used in reference to any type of communications, whether textual, visual, or auditory.

[16] This tool is used in information retrieval to identify documents where a desired search term occurs with relatively high frequency.

software can also be used simply to process text using customized wordlists selected by the user.) Examples of studies outside of accounting and finance that use Diction include "Media, terrorism, and emotionality: Emotional differences in media content and public reactions to the September 11th terrorist attacks" (Cho et al. 2003) and "The power of leading subtly: Alan Greenspan, rhetorical leadership, and monetary policy" (Bligh and Hess 2007). The GI wordlists have been used for more than four decades and were developed in the domain of social psychology. Examples of applications using the GI lists include "Prosody and lexical accuracy in flat affect schizophrenia" (Alpert et at. 2000) and "Some characteristics of genuine versus simulated suicide notes" (Ogilvie et al. 1966). As is evident by the titles of these works, the GI wordlists like Diction, have been applied in a wide variety of settings.

There is a growing interest in applications of computational linguistics to disclosure in accounting and finance, and many of these studies use the positive and negative wordlists from either Diction or GI. Using the Diction positive and negative wordlists, Davis et al. (2007) find a relation between the market reaction to earnings announcements and the unexpected tone of the announcement. Rogers et al. (2009) also use the Diction positive and negative wordlists and find that the tone of sued firms' disclosures are more positive than disclosures by a matched sample of non-sued firms, and suggest that the use of more positive language increases the likelihood of being sued. Both studies use equal weighting: Davis et al. (2007) define *POSITIVE* and *NEGATIVE* as the percentage of positive and negative words, respectively, and Rogers et al. (2009) measure tone as the difference between the counts of positive and negative words, scaled by total words.

Using the GI positive and negative wordlists and equal weighting, Kothari et al. (2009) show that negative tone in management's disclosures is significantly associated with the firm's stock return volatility and analysts' forecast error dispersion though not with a firm's cost of capital. In another study using the GI positive and negative wordlists and equal weighting, Tetlock et al. (2008) find that a firm's future earnings

and future stock returns are related to the tone of news stories about the firm and that the relationship is even more significant for stories that mention the word "earnings."[17]

As an alternative to using wordlists and frequency counts to measure tone, some researchers have used classifier algorithms that entail the following general steps: assigning a training data set into a particular classification using some reliable mechanism (typically manual coding)[18] and then using the results of the manually-coded data as input to a classifier algorithm that then automatically classifies a larger data set into different categories. For example, Antweiler and Frank (2004) manually classify a training data set of 1,000 internet bulletin board messages and use two classification algorithms (the Naïve Bayes classifier and the Support Vector Machine algorithm) to classify a sample of 1.6 million messages into one of three categories: buy, hold, or sell. Das and Chen (2007) combine classifier algorithms with a frequency count tone measure to classify 145 thousand stock message board postings for tech-sector stocks into buy, hold, or sell categories. These researchers manually classify a training set of 1,000 postings, use five different classifier algorithms to indicate how the posting should be classified, choose the classification based on the majority 'vote' of the five classifiers, and refine the classification methodology by also using a tone score (based on the GI wordlist) to eliminate ambiguous messages. Li (2010) manually classifies 30,000 forward-looking sentences from 10-Ks with the help of 15 research assistants (all of whom had completed at least an intermediate accounting course) and then uses the Naïve Bayes classifier to classify a sample of 13 million forward-looking sentences as positive, negative, neutral or uncertain. Li (2010) then uses the average tone of all the classified forward-looking sentences within a management discussion and analysis (MD&A) to classify the tone of that MD&A as one of three categories: positive, neutral or negative (which also includes uncertain).

[17] Tetlock et al. (2008) define the negative words variable as the percentage of negative words in news stories from 30 to 3 trading days prior to an earnings announcement, but also state that similar but weaker results are found using alternative definitions that incorporate positive words.

[18] Instead of manual coding, Balakrishnan et al. (2010) use one year of their data as a training set and assign the documents to one of three categories (out-performing, average and under-performing) based on the company's actual market performance over the year following the time of their training document.

This study focuses on wordlist-based tone measures for several reasons. First, the initial step in classifier algorithms typically involves human coding which – apart from being extremely labor intensive – poses potential problems of inter-rater reliability both within a given study and across studies. For example, Das and Chen (2007) report that the human coders agreed on the classification of only 72 percent of the messages in their training sample. While wordlist-based tone measures may be less nuanced than manually assessed tone categories, creation of those measures is far less labor intensive and inter-rater reliability is not an issue. Second, replication is far more difficult with classifier algorithms than with wordlist-based tone measurement. In addition to potential problems of inter-rater reliability of human coding across studies, an algorithm classifier, by definition, produces classifications that are specific to the particular training data set. Different classifier algorithms also incorporate different research design choices such as the validation process (see e.g., Li 2010 for a discussion of validation alternatives using the Naïve Bayes method), the assumed misclassification costs and prior probability of each classification category (see e.g., Henry 2006 for an example of these assumptions in the CART classification algorithm), and the minimum frequency count for inclusion in the classifier vocabulary restrictions (see e.g., Antweiler and Frank for an example of an assumed setting for the "prune-vocab-by-infogain" option when implementing the Naïve Bayes method in the Rainbow software package. Given these difficulties, two studies using the same classification algorithm are extremely likely to result in different classifications of the same document. In contrast, two researchers applying the same wordlist and tone definition to the same document will obtain identical tone measures.

2.1 General versus domain-specific wordlists

Though researchers have had some success using positive and negative wordlists from GI and Diction, one potential concern is that the language used in accounting disclosures tends to be somewhat specialized, essentially a sub-language pertaining to a specific domain of discourse. Consequently, those broad wordlists likely lack predictive power in a capital markets setting. The GI wordlist contains many words that are likely irrelevant to financial communication. For example, the GI positive wordlist includes the words "witty," "wonderful," "wondrous," "woo," "zest," "delight," "dazzle," and "cupid" to name only a

few of many words that appear irrelevant to financial communication. As another example, the GI positive word list omits the words "record" and "strong" both of which are used in earnings announcements to connote positive aspects of performance (e.g., Henry 2008, 11-12). Further, although the GI negative wordlist includes the word "decrease," the GI positive wordlist does not include the word "increase." Finally, numerous words categorized as negative or positive in the non-domain specific wordlists based on one meaning of the word; however, these words often have different meanings in financial disclosure. For example, GI categorizes the word "beat" as negative, but in earnings press releases, this word typically implies that companies have exceeded expectations.

An alternative to using general wordlists (e.g., Diction and GI) is to utilize a domain-specific wordlist. Within the body of capital markets research, one of the tone scores we examine – which we refer to as the FD score – uses a wordlist that is specific to the domain of financial communication (Henry 2006, 2008). Henry (2006, 2008) constructs a wordlist intended specifically for financial disclosure and demonstrates its predictive validity in the context of earnings announcements. Henry's (2006) main findings are that word count measures of qualitative information in earnings announcements (one component of which is positive or negative tone) improve prediction of firms' returns following the announcement. Henry's (2008) main findings are the tone of qualitative information in earnings press release has a positive, nonlinear relation with abnormal returns during the three-day event window around the earnings announcement. Henry (2008) defines tone as (*POSITIVE – NEGATIVE)/ (POSITIVE + NEGATIVE*) where *POSITIVE* is the frequency of occurrences in an earnings announcement of words on the FD positive wordlist and *NEGATIVE* is the frequency count of occurrences of words on the FD negative wordlist. Another domain-specific wordlist, developed in Loughran and McDonald (2011), has been used to develop a tone measure with higher explanatory power than a tone measure using the GI wordlist.

In spite of the advantages of domain-specific wordlists for assessing tone, most of the early studies in accounting and finance use general wordlists. In the following section, we describe tests to show the empirical advantages of using a domain-specific wordlist for measuring the tone of financial disclosure. Our objective is to measure the relative effectiveness of these scores in capturing the underlying tone of

financial disclosure, which in this case is the earnings announcement. One research design issue is how to identify the "true" underlying tone of the disclosure in order to compare the relative efficacy of the alternative measures we consider. Rather than attempting to measure the "true" tone, we consider the relative impact that the various tone measures have on stock returns around the earnings announcement – i.e., their relative predictive validity. Past research has shown that the tone of earnings announcement disclosure is related to event returns. Therefore, we can assume that if one score is more closely associated with stock returns (e.g., the coefficient is greater in magnitude), it is a better measure of the underlying tone. Further, we can assume that if one score better explains the variation of event window stock returns (e.g., the R-squared of the event-window regression is higher), it is a better measure of the underlying tone.

Our approach is appealing because it shows the relative benefits of the alternative measures using actual data (rather than through simulation), in a setting where tone has been shown to be related to stock returns. However, the validity of our tests depends on the assumption that one measure does not suffer any more or less from a correlated omitted variables problem than the others. For example, one wordlist could contain a word that is closely related to risk but is not included in the other wordlists. Thus, one wordlist might be more highly correlated with risk, which is the underlying cause of the returns (as opposed to tone). If this is the case, then the score shown to be more highly correlated with stock returns actually suffers most from a correlated omitted variables problem (causing Type I errors) and is not necessarily better at detecting the underlying tone. Though we review the wordlists for such potential problems, we cannot completely rule out this alternative interpretation.

2.2 Weighting

A second choice researchers must make is the weighting of words in a wordlist. The majority of research in accounting and finance employ equal weighting, where each occurrence of a word in the wordlists is weighted the same. In a related paper, Loughran and McDonald (2009), recommend the addition of document frequency-weighting methodologies when creating word count measures of

qualitative information in financial disclosure.[19] The authors use inverse document frequency weightings to produce word-frequency-inverse-document-frequency measures (*wf-idf*),[20] widely used in the information retrieval domain (e.g., Google and other search applications), and show empirically that after applying this weighting method, the GI wordlist is at least as powerful as their own domain-specific wordlist (Loughran and McDonald 2009).

The *wf-idf* method favors words that occur in fewer documents within a sample because those words improve the ability to rank documents according to their relevance to a particular search and thus to obtain more precise search results. The *idf* weight is defined as $\log(N/n_j)$ where N is the total number of documents in the sample and n_j is the number of documents that contain the search word. If a search word appears in many documents, it contributes little to ranking documents' relevancy to the search. If a search word appears in every document in the sample, its IDF weight would be log(1)=0; it would contribute nothing to ranking documents.

In the context of information retrieval tasks, where the objective is to sift through millions of documents and rank order those documents by their relevance to the search criteria, a document-weighting methodology such as *wf-idf* is clearly advantageous. The word-frequency measure *wf* serves to capture the content of a particular document, and the *idf* weighting serves to rank order the documents by their relevance to the search term. In tasks solely involving content analysis, for example when a document of interest has already been identified, the *idf* weighting is arguably unnecessary.

Studies that focus on analyzing the content of documents that have already been identified obviously do not need to employ tools aimed at locating the documents. Comparable research in other disciplines (e. g, political communication, psychology, etc.) focused on the analysis of content of pre-identified documents do not add *idf* weightings, and we caution against the use of *idf* weightings in the

[19] While Loughran and McDonald (2009) also examine a domain-specific wordlist, similar to us, the key difference with our paper is that they advocate the sample-dependent, *wf-idf* methodology, while we caution against its use. Other differences include: our use of earnings announcements rather than 10-K filings, which we think better reflects discretionary disclosure; our use of event-study methodology; our demonstration of relative predictive ability of the FD wordlist at smaller sample sizes; and our use of a more parsimonious wordlist that has been validated in prior capital markets research.

[20] See Manning and Schutze's (2002) "Topics in Information Retrieval" for an explanation of *idf* weightings.

context of analyzing financial disclosure. The commonality of a word across a sample generally does not alter the information of a given document and thus should not impact its weighting. Consider, for example, a sample consisting of three documents, with each document containing a total of 100 words. The word "less," appears in Document 1 (D1) ten times and does not appear in D2 or D3. The word "lower," appears in D2 ten times and D3 one time, but does not appear in D1. Further, these are the only negative words from the wordlist that appear in the documents. Under an equal weighting based solely on word frequency, which we advocate, the negative tone scores with a simple scaling would be 0.10, 0.10, and 0.01 for D1, D2, and D3, respectively. However the *idf* weight for the word "less" is 1.1 (i.e., ln(3/1)) and the *idf* weight for the word "lower" is 0.41 (i.e., ln(3/2)). Combining the word frequency weight *wf* and *idf* weights would give *wf-idf* measures of 0.1100, 0.0410, and 0.0041, respectively.[21] Even though D1 and D2 contain the same number of negative words, *idf* weights would score D1 as being nearly three times more "negative" than D2, simply because the word "lower" is also contained one time in another document in the sample. With *wf-idf* weightings, the tone score of any given document would always depend on the contents of every other document in the sample.

Loughren and McDonald (2011) argue that an advantage of *wf-idf* weighting is that it serves to correct for misclassified words in a wordlist. For example, the GI wordlist misclassifies –*vis a vis* financial disclosure – the word "division" as negative. If this word appears in most documents in the sample, it will be down-weighted significantly so that its impact on the tone measures will become negligible. Although *wf-idf* weighting will mitigate the impact of misclassification for very common words, it will actually exacerbate the impact of misclassified words that occur in a small number of documents. Additionally, *wf-idf* weighting decreases the impact of words that are correctly classified but also appear at least once in a large number of documents across the sample. Consequently, we argue that the best way to eliminate the problem of misclassification is simply to correct these misclassifications in the wordlists (e.g., remove the

[21] See Appendix A for an example computation *wf-idf* and more detailed examples of problems associated with the application of *wf-idf* to tone scores.

word "division" from a wordlist designed to capture negative financial disclosure). Results in Sections 4 and 5 are presented using both equal weighting and *idf*-weighting for each tone measure.

In summary, *wf-idf* weighting is an effective information retrieval methodology but is less useful for measuring tone, and potentially distorts the relative tone scores one would obtain under equal weighting. Moreover, *wf-idf* weightings are, by design, sample dependent. The advantages of equal weighting are that it is simple, transparent, and amenable to replication. [22]

3.0 Sample Selection and Definition of Tone Measures

Beginning in 2003, publicly listed companies are required to furnish earnings releases and similar materials to the SEC on Form 8-K (SEC 2003). We obtain earnings announcements from the SEC FTP site between 2004 and 2006. We select all 8-K filings that include an earnings release identified by Item Code 2.02. There are 47,376 such filings between 2004 and 2006.[23] We then merge this data with Compustat and CRSP, which reduces the sample to 34,135 firm years. After deleting observations with missing variables, including all tone measures, the sample includes 29,712 firm years.

We process the 8-K filings using Perl scripts. We first strip out the earning release from any other items included with the 8-K filings. We also remove any HTML or SGML tags. Next, the Perl script conducts a frequency count for words in each of the four sets of word lists (FD, DICTION, GI, and LM). We "preprocessed" the GI wordlist to account for word "stems." For example, the entry "decrease" on the GI list is a stem (the primary lexical root of the word); it is the singular form of the noun and the infinitive form of the verb. A frequency count of the word, in all its forms, can be obtained either by instructing the text-processing program to include all words that contain the stem or, alternatively, by expanding the list itself to include all forms of the word – "decrease, decreases, decreased, decreasing." We use the second

[22] As Tetlock et al. (2008) write about their word count measures, which employ proportional weighting: they "are parsimonious, objective, replicable, and transparent. At this early stage in research on qualitative information, these four attributes are particularly important, and give word count measures a reasonable chance of becoming widely adopted in finance."

[23] This is not an all-inclusive list of earnings press releases. One shortcoming of the SEC filings is that the Item codes are unaudited and a large number of firms miscode their 8-K filings. We only include firms that properly classified their 8-K filing as an earnings release. Note also that the new 8-K coding did not take effect until August, 2004, so our sample effectively starts in August 2004.

approach because both the FD and Diction lists include complete words rather than stems.[24] The wordlist used to create the FD score includes a total of 93 negative words and 188 positive words; all focused on the domain of earnings announcements. The wordlist from Diction used in accounting research to measure positive and negative includes 914 negative words and 697 positive words. The wordlists from GI contain 3,699 negative words and 2,557 positive words. The negative LM wordlist contains 2,337 words, and the positive LM wordlist contains 353 positive words.[25]

Once the frequency counts are complete, we compute tone scores using each of the word lists. For comparability to prior literature, we examine a "net tone" score as well as "positivity," and "negativity" scores. Each of the measures we consider are named and described as follows:

Net Tone Scores

FD_SCORE =(POSITIVE – NEGATIVE)/ (POSITIVE + NEGATIVE) where *POSITIVE* and *NEGATIVE* refer to the word count frequency based on the positive and negative words in the FD word list;

DICTION_SCORE =(*POSITIVE – NEGATIVE)/ (POSITIVE + NEGATIVE)* where *POSITIVE* and *NEGATIVE* refer to the word count frequency based on the positive and negative words in the DICTION word list;

GI_SCORE =(POSITIVE – NEGATIVE)/ (POSITIVE + NEGATIVE) where *POSITIVE* and *NEGATIVE* refer to the word count frequency based on the positive and negative words in the GI word list;

LM_SCORE =(POSITIVE – NEGATIVE)/ (POSITIVE + NEGATIVE) where *POSITIVE* and *NEGATIVE* refer to the word count frequency based on the positive and negative words in the LM word list.

Positivity Scores

POS_FD =POSITIVE/NUM_WORDS where *POSITIVE* is the word count frequency based on the positive FD wordlist, and *NUM_WORDS* is the total number of words in the document being analyzed;

[24] One reason that the FD wordlist does not use stemming is that not all forms of a word are necessarily equally indicative of positive or negative information, particularly in a specific domain of discourse. For example, in earnings announcements, the word "record" used as an adjective, as in "record sales" is positive, while the word "recorded" would typically be used as the past tense of the verb without positive or negative implications.

[25] Loughran and McDonald (2011) use a relatively long list of words arguing that a longer list of words will make it more difficult for a manager to systematically avoid using the words. Our view is that (a) the role of a measure of qualitative information is to best capture the tone of that information, not to shape future managerial behavior, and (b) parsimony is a generally desirable attribute of tone measures.

POS_DICTION =POSITIVE/ NUM_WORDS where *POSITIVE* is the word count frequency based on the positive DICTION wordlist, and *NUM_WORDS* is the total number of words in the document being analyzed;

POS_GI=POSITIVE/ NUM_WORDS where *POSITIVE* is the word count frequency based on the positive GI wordlist, and *NUM_WORDS* is the total number of words in the document being analyzed

POS_LM=POSITIVE/ NUM_WORDS where *POSITIVE* is the word count frequency based on the positive LM wordlist, and *NUM_WORDS* is the total number of words in the document being analyzed.

Negativity Scores

NEG_FD =NEGATIVE/ NUM_WORDS where *NEGATIVE* is the word count frequency based on the negative FD wordlist, and *NUM_WORDS* is the total number of words in the document being analyzed;

NEG_DICTION=NEGATIVE/ NUM_WORDS where *NEGATIVE* is the word count frequency based on the negative DICTION wordlist, and *NUM_WORDS* is the total number of words in the document being analyzed;

NEG_GI=NEGATIVE/ NUM_WORDS where *NEGATIVE* is the word count frequency based on the negative GI wordlist, and *NUM_WORDS* is the total number of words in the document being analyzed;

NEG_LM=NEGATIVE/ NUM_WORDS where *NEGATIVE* is the word count frequency based on the negative LM wordlist, and *NUM_WORDS* is the total number of words in the document being analyzed. For the net tone scores, a purely positive press release would have a tone score of one, a purely negative press release would have a tone score of negative one, and a perfectly neutral press release would have a tone score of zero. For each of the positivity and negativity scores, we also create variables as the natural logarithm of the frequency count of the positive words and negative words in the document. In addition, we create *idf*-weighted measures following Manning and Schutze (2002). Specifically, when the word frequency weight is calculated as the log of the word frequency count and the idf-weighting is the inverse document frequency, the *wf-idf* measure can be formed as shown below (from Manning and Schutze, 2002, 543).

$$wf\text{-}idf = \begin{cases} (1 + \log(wf_{i,j}))\log(N/n_j) & \text{if } wf_{i,j} >= 1 \\ 0 & \text{if } wf_{i,j} = 0 \end{cases} \quad (A1)$$

For each of the alternative wordlists, the idf-weighted positive tone score and idf-weighted negative tone score are scaled by the total wordcount of the document. The overall idf-weighted tone Score is equal to (idf-weighted positive tone - idf-weighted negative tone)/ (idf-weighted positive tone + idf-weighted negative tone).

4.0 Results: Descriptive Statistics and Tests of Market Reaction to Tone of Earnings Announcements

Descriptive statistics for the tone scores of our sample of 29,712 earnings announcements are shown in Table 2. The mean and median values of both *FD_SCORE* (mean=0.429 and median=0.462) and *GI_SCORE* (mean=0.334 and median=0.337), are significantly higher (i.e. more positive) than the mean and median values of *DICTION_SCORE* (mean=0.044 and median=0.043) or LM_Score (mean = -0.037 and median = -0.067).[26] Prior research on the tone of financial disclosure indicates that voluntary disclosure tends to be positive. For example, Abrahamson and Amir (1996, 1163) explain their decision to focus only on negative statements in the presidents' letters appearing in annual reports:

> A quick look at a number of president's letters reveals that they are 'sugar coated.' That is, they are full of positive statements. Coding such positive statements, most of the sugar-coating turns out to be irrelevant and ritualistic (our employees are happy, our sales went up, etc.). It would be a waste of effort to sift through this large number of meaningless statements to find the important ones.

Similarly, Rutherford (2005) shows that language in annual report narratives is biased toward the positive. Thus, based on prior research of other financial narratives, we would expect that the average values of the

[26] The likely reason that the mean and median of the tone score using the LM wordlists are negative is that the LM wordlist contains nearly 7 times more negative words than positive. In contrast, for the other three wordlists, the number of negative words is much closer to the number of positive words. Specifically, the ratio of Max(negative words, positive words) to Min(negative words, positive words) is 2 or below for the FD, Diction and GI wordlists.

tone score for earnings press release to be greater than zero. The consistency between the expected bias toward positivity and the higher levels of *FD_SCORE* and *GI_SCORE* score suggest that these two measures have more face validity than *DICTION_SCORE* or *LM_SCORE* in capturing the tone of voluntary financial disclosure.

All of the tone scores have a maximum value of 1.00, indicating a purely positive tone. For *FD_SCORE* , *DICTION_SCORE* and *LM_SCORE*, the minimum value is minus 1.00, but the *GI_SCORE* has a minimum of -0.302. Table 2 also presents descriptive statistics for the decomposed positive and negative tone scores derived from each wordlist. The higher values of *NEG_GI* and *POS_GI* compared to f the other three wordlists can be explained by the fact that the GI wordlist is simply longer than the other three lists, increasing the probability that at least one of the words on its list will occur in any given press release.

Table 3 reports the correlation between each of the tone measures and the following: cumulative abnormal returns around the earnings announcement date *(CAR),* unexpected earnings *(UE)*, the size of the announcing firm based on total assets *(SIZE)*, and the length of the earnings announcement measured as the word count *(NUM_WORDS)*.

For the separate positive and negative tone measures, all signs are as expected. The negative tone measures (*NEG_FD, NEG_DICTION, NEG_GI,* and *NEG_LM*) are all negatively correlated with *CAR*, and the positive tone measures (*POS_FD, POS_DICTION*, *POS_GI*, and *POS_LM*) are all positively correlated with *CAR*. The magnitude of the correlation coefficients varies across the tone measures. Because the magnitude of the separate positive and negative tone measures depend on the number of words in the wordlists, the relative magnitude cannot be compared. In contrast, the composite tone scores use the same scaling, so the magnitude of the coefficients can be compared. As shown, the *CAR* is more closely correlated with the *FD_SCORE* than with any of the other score measures.

4.1 Empirical Tests of the Relation between Market Reaction and Alternative Measures of the Tone of Earnings Announcements

We employ a short-window event study methodology to examine the relation between market reaction to earnings announcements and the tone of the qualitative information in the announcements, measuring tone alternately using the domain-specific and non-domain specific wordlists. Using past research as our starting point, we begin with the following base regression:

$$CAR = \alpha + \beta_1 UE + \beta_2 SIZE + \beta_3 X\_SCORE + \varepsilon \qquad (1)$$

Where

CAR = cumulative abnormal returns from day t-1 to day t+1 around earnings announcement date;

UE = unexpected earnings per share (Compustat variable *EPSPXQ*) in quarter *t* - earnings per share in quarter *t*-4, scaled by price in *t*-4;

SIZE = Log of total assets in period *t* (Compustat variable *AT*);

X_SCORE= is the net tone score, alternately *FD_SCORE, DICTION_SCORE, GI_SCORE,* and *LM_SCORE*, of the current earnings press release.

We estimate this regression for each net tone score and, since all tone score measures are of the same scale (ranging from a possible maximum of 1 and minimum of -1), we consider both the magnitude of the coefficient and the significance levels to assess the relative power of the alternative measures.

Table 4 presents the results.[27] Consistent with the univariate results, the coefficients estimated for each of the tone scores are positive and significant. The magnitude of the coefficient on the domain specific tone score, *FD_SCORE,* is greater than the coefficient on *DICTION_SCORE, GI_SCORE* or *LM_SCORE*. Results of Vuong tests (1989) indicate that the models incorporating the FD score and the LM score have greater predictive ability than either Diction score or GI score, while neither of the latter two models dominates.

We next add a *LOSS* indicator (*LOSS*=1 if the firm reports a loss in the current quarter) to the base regression for two reasons. First, *LOSS* controls for the differential stock price reaction to earnings announcements by loss-making firms (Hayn 1995). Second, both of the general wordlists, GI and DICTION, and the LM wordlist include the word "loss" as a negative word, but the FD negative wordlist

[27] Reported results include year fixed effects in regressions, with robust standard errors clustered by firm. Untabulated results using year and firm fixed effects yield the same conclusions.

does not; therefore, so addition of the *LOSS* variable enables a more specific evaluation of the overlap in information between the accounting data (i.e. earnings less than zero) and the qualitative information captured by the wordlists used to measure tone.

Controlling for loss-making firms reduces the magnitude of the coefficient on *DICTION_SCORE* by roughly 46 percent (from 0.013 to 0.007), on *GI_SCORE* by around 42 percent (from 0.024 to 0.014), and on LM_SCORE by around 28 percent (from 0.018 to 0.013), but reduces the magnitude of the coefficient on *FD_ SCORE* by only 15 percent, as shown in Table 5 compared with Table 4. Further, the economic significance of the FD score is nearly three times that of either the Diction score or the GI score, and roughly 30 percent higher than that of the LM score. Specifically, a one standard deviation change in the FD score increases *CAR* by 0.0825 standard deviations versus an increase of 0.0313, 0.0299, and 0.0622 for Diction, GI, and LM scores, respectively. This suggests that the occurrence of an accounting loss contains a significant amount of overlapping information, particularly with the two non-domain specific tone measures but less so for *FD_SCORE*. In other words, because the information that a firm reported earnings less than zero is captured by the quantitative disclosure, the tone measures employing the word "loss" (*DICTION, GI,* and *LM*) have less incremental informativeness beyond financial-statement data compared to a tone measure that excludes financial statement terms (*FD*).

In untabulated results, we repeat the analysis in Tables 4 and 5 using the change in the tone scores from the previous period rather than the level of the tone scores. Key results hold in these analyses: regressions using the *FD_SCORE* continue to have greater explanatory power than those using the other tone measures.

4.2 Tone Scores Decomposed Into Positive and Negative Tone

In addition to testing the impact of net tone, we consider positivity and negativity separately. By doing so, we can compare our results to past research that focuses on either negativity or positivity, and on the possible asymmetric market reaction. As noted, Tetlock (2007) and Tetlock et al. (2008) find a stronger relation between negativity and stock returns than between positivity and stock returns. We estimate the following regression using all four wordlists as follows:

$$CAR = \alpha + \beta_1 UE + \beta_2 SIZE + \beta_3 LOSS + \beta_4 NEG\_X + \beta_5 POS\_X + \varepsilon \quad (2)$$

Where, all variables are as defined above; and

NEG_X=NEG_FD, NEG_DICTION, NEG_GI, or NEG_LM, depending the score being tested;

POS_X= POS_FD, POS_DICTION, POS_GI, or POS_LM, depending the score being tested.

Table 6 presents results of the regressions with separate measures for positive and negative tone. We report results including the scaled measures described above and measures calculated as the natural logarithm of one plus the positive word count and negative word count, respectively. We obtain similar results with the transformed variables and, therefore, only report results with untransformed values for brevity. As shown in Table 6, both NEG_X, and POS_EX are in the expected direction for all measures (FD, DC, GI, and LM). However, the coefficient on NEG_X is insignificant for Diction. Both domain-specific wordlists (FD, LM) outperform the general wordlists (Diction and GI). For the GI wordlist, the positive tone measures have a lower level of statistical significance (p-values < 0.05) compared to that of the negative tone measures (p-values <0.001), consistent with findings in Tetlock (2007) and Tetlock et al. (2008), which use the GI wordlist.

Results also indicate that even for the FD scores, the market reaction to negative tone is greater than market reaction to positive tone. A one standard deviation change in negative tone is associated with a 0.071 standard deviation change in *CAR*, but a one standard deviation change in positive tone is associated with only a 0.057 standard deviation change in *CAR*. This is consistent with investors placing less weight on positive qualitative disclosure than on negative (Hoskin et al. 2002; Hutton et al. 2003; Tetlock 2007; Tetlock et al. 2008).

4.3 Tone Scores with IDF Weightings

The results described above use equal weighted tone scores. To examine the incremental explanatory power provided by using *idf* weightings, Table 7 presents the regressions for each unweighted tone measure side by side with the *idf*-weighted measures. As shown, using *idf*-weighted measures improves the explanatory power in each specification except for Diction, where the change is slightly

negative. Results (untabulated) of Vuong's test indicate the increase in R-square from using *idf* weights is statistically significant only for the FD measure.

5.0 Empirical Tests of the Relative Power of Alternative Tone Measures

In the previous section, we show that, in a regression of abnormal returns on unexpected earnings and alternative measures of tone, the coefficients on the FD tone scores are larger in magnitude than the corresponding coefficients on the GI and Diction tone scores. In addition, the explanatory power of regressions including the FD score exceeds that of the regressions using either of the other scores. We infer from this that the FD scores are superior to the Diction, GI, and LM scores in capturing management's tone in the press release. In this section, we extend our analysis by testing the relative statistical power of the four measures to reject the null (of no incremental information content in tone) when sample sizes are relatively small.

As in the previous section, we employ a short-window event study methodology to examine the relation between market reaction to earnings announcements and the tone of the qualitative information in the announcements power of the alternative tone measures, first using equal weighting and then using *idf*-weighting. We explore the relative power of the four tone scores using randomly selected sub-samples of varying sizes ranging from 50 observations to 2,000 observations. We select 250 sub-samples (with replacement) of each size, and estimate the regression equations above on each. Figure 1A plots the rejection rates of the coefficients generated from regressions of equation (1) using the four unweighted tone scores. Regardless of sample size, the coefficient on *FD_SCORE* is greater than the other three scores. Figure 1A shows that the rejection rates converge to almost 100% as the sample size increases to 2,000. The *LM_SCORE* also increases significantly as the sample size increases to 2,000 (roughly 79%), but the rejection rates for the generalized tone scores remain quite low. This suggests that the power of generalized wordlists to detect tone is relatively low in this financial disclosure setting.

Figure 1B, presents results using *idf*-weighted tone scores. For the idf-weighted measures, a separate IDF weighting is calculated for each of the 1,500 different subsamples, simulating the process that

a researcher would take. The rejection rate for the *idf*-weighted FD and LM scores is higher for each of the sample sizes but the improvement is quite modest and decreases as the sample size increases. For the Diction and GI scores, however, *idf*-weighting improves explanatory power at certain sample sizes and reduces explanatory power at other sample sizes, demonstrating the corpus-dependency of *idf*-weightings.

Figure 2 shows the comparison of rejection rates for the tone scores decomposed into positive and negative. The patterns in Figures 2A and 2B are consistent with those in Figure 1. The rejection rates for *NEG_FD* increase from 12% percent for sample sizes of 50 to 91.6% percent for sample sizes of 2,000. These results are far better than those for any of the other measures. In the case of positive tone, POS_FD consistently outperforms the other measures but POS_LM comes fairly close to POS_FD when the sample size is 2,000. The rejection rate for POS_FD is 74.8% when n=2000, compared to 60.4% for POS_LM. The rejection rates are much lower for POS_DC and POS_GI, with rates of 22.4% and 7.6%, respectively. These results lead to a similar conclusion, that domain-specific wordlists, particularly the FD wordlist, are more powerful in the context of financial disclosure.

6.0 Analysis of the Key Reasons for the Differences in Tone Scores

To understand the reasons for the differences in the scores, we identify and analyze earnings announcements for which the different wordlists yield the greatest measurement differences (i.e., *FD_SCORE* is very different from *GI_SCORE*). For each press release, we compute a score divergence as *FD_SCORE* minus *GI_SCORE*. If the score divergence is a negative (positive) number, the FD wordlist measures the tone of the press release as more negative (positive) than the GI wordlist. We identify the top 5 percent press releases (1,485 releases), ranked by divergent scores – one group of press releases measured as more negative using the FD wordlist and one group of press releases measured as more positive using the FD wordlist. The most divergent releases are then examined by group.

A press release's FD tone score can be more negative than its GI tone score because it more frequently uses words that are included in (excluded from) the FD negative wordlist (the GI negative wordlist) and/or because it more frequently uses words excluded from (included in) the FD positive wordlist

(the GI positive wordlist). For the group of press releases measured as much more negative using the FD wordlist, Panel A of Table 8 lists the most frequently occurring words appearing in the negative FD wordlist but not in the negative GI wordlist, and Panel B of Table 8 lists the most frequently occurring words appearing in the positive GI wordlist, but not in the FD wordlist.[28] Across both panels, the most word that occurs in the largest number of this group of press releases is the word "share" which the GI list includes as a positive word. ("Shares" is also among the most commonly occurring words. See the above for the explanation why stemming is not used in this study.)

Used in certain ways – for example, as a verb meaning allowing someone else to use something that one possesses (e.g., I share my thoughts with you) or having in common (e.g., we share a love of sports) – the word "share" has positive connotations. In the domain of financial disclosure, however, it can be expected that the word "share" refers to a certificate of ownership of a company and thus has neither positive nor negative connotations.

For the group of press releases measured as much more positive using the FD wordlist, Panel A of Table 9 lists the most frequently occurring words appearing in the positive FD wordlist but not in the positive GI wordlist, and Panel B of Table 9 lists the most frequently occurring words appearing in the negative GI wordlist, but not in the FD wordlist. Across both panels, the most frequently occurring words are "increase" and "increased" which the FD list includes as a positive word but the GI wordlist does not.

Appendix B presents an excerpt from an earnings press release where the FD tone score was much more positive than the GI tone score. The issuing firm is Monster Worldwide, the online advertising company. The divergence in tone scores arises from the GI wordlist's inclusion of the words "monster" and "division" as negative words, and the GI wordlist's exclusion of words such as "increase", "strong", and "growth." This analysis supports our conjecture and statistical analysis that researchers are on much firmer ground if they use a domain-specific wordlist (e.g., *FD_SCORE*). General wordlists are much more likely to

[28] We limit the list to the most frequently occurring, non-overlapping tone words because, as is generally the case in computational linguistics, the distribution of occurrences of words is highly skewed (i.e., most words are rare).

include words with meanings that do not match the context of financial disclosure, or exclude words would be viewed as positive or negative in a financial disclosure.

7.0 Conclusion

In this study we evaluate methods used in accounting and finance to measure the tone of financial disclosures. A growing stream of capital markets research examines how stock prices reflect the tone of language used in financial disclosures, also referred to as the qualitative information, controlling for concurrently-disclosed quantitative information. We consider measures of tone based on four alternative wordlists used in capital markets research: 1) a wordlist developed in Henry (2006, 2008) specifically for use in the domain of financial disclosure (*FD_SCORE*); 2) a wordlist from Diction software (*DICTION_SCORE*); 3) a wordlist from the General Inquirer (GI) program (*GI_SCORE*); and 4) a wordlist developed in Loughran and McDonald (2011).

Using a sample of 29,712 earnings press releases in a short-window event study, we find that the FD tone score is more powerful than those constructed from the Diction, GI, or LM wordlist. We also find that the decomposed FD scores and LM scores (both positive and negative) are incrementally informative in regressions of returns on unexpected earnings and tone scores, unlike the other two scores. However, similar to Tetlock (2007), we do find a smaller economic impact of positive tone compared to negative, even using the FD scores. This is consistent with investors placing less weight on positive qualitative disclosure than on negative (Hoskin et al. 2002; Hutton et al. 2003; Tetlock 2007; Tetlock et al. 2008).

In addition to our large sample tests, our detailed analysis of the press releases where the tone scores diverge the most helps us to understand why the FD tone score dominates the other measures. We identify frequently-occurring words in the positive GI wordlist such as “share” and “company” whose domain-specific meaning – in our judgment – is not indicative of positive qualitative information and, to a lesser extent words in the negative GI wordlist such as “press” and “tax” whose domain-specific meaning – in our judgment – is not indicative of negative qualitative information. These findings support Engelberg’s

(2008) speculation that the lack of a relation between stock returns and positive qualitative information is due to error in measuring the qualitative information when tone scores are based on the GI wordlist.

Finally, in contrast to Loughren and McDonald (2009), we argue that the addition of a sample-dependent *idf* weighting to a document-specific word frequency *wf* weighting, which is popular for use in search algorithms, does not logically translate to measuring tone. The most appropriate method for weighting in the context of measuring the tone of a given disclosure is document-specific equal weighting, because it is not sample-dependent. Our analysis indicates that *idf*-weighting can increase the power of tone measures in short-window event studies for some wordlists, for some samples. The variability in the improvement demonstrates the corpus-dependency of *idf*-weightings. In contrast with Loughran and McDonald (2011), we caution against using *idf*-weightings because they introduce impediments to replication. Although using an *idf* weighting corrects for mistakenly classified words – the chief weakness of using a wordlist that was not developed for use in the domain of financial disclosure – we argue that a more transparent and simpler method is simply to use a wordlist that does not contain such mistakes.

Overall, our findings suggest that capital markets researchers aiming to measure the qualitative information in financial disclosure can significantly increase the power and simplicity of their tests by using a domain-specific wordlist (e.g., FD or LM) with equal weighting. Future capital markets research on this topic can address variations in the wordlists best suited to capturing qualitative information in various types of financial narrative (e.g., companies' earnings announcements, management discussion and analysis, CEO letters to shareholders in annual reports, analysts' reports, and financial journalists' discussion of company performance) and changes in those wordlists over time.

References

Abrahamson, E., and E. Amir. "The Information Content of the President's Letter to Shareholders." *Journal of Business, Finance & Accounting* 23 (1996): 1157-1182.

Ajinkya, B.; S. Bhojraj; and P. Sengupta. "The Association Between Outside Directors, Institutional Investors and the Properties of Management Earnings Forecasts." *Journal of Accounting Research* 43(2005): 343-376.

Alpert, M.; Rosenberg, S.; Pouget, E.; and R. Shaw. Prosody and lexical accuracy in flat affect schizophrenia. Psychiatry Research 97 (2000): 107-118.

Antweiler, W., and M. Z. Frank. "Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards." *Journal of Finance* 59 (2004): 1259–94.

Balakrishnan , R.; X.Y. Qiu; and P. Srinivasan. "On the Predictive Ability of Narrative Disclosures in Annual Reports." *European Journal of Operational Research* 202 (2010): 789-801.

Berelson, B. *Content Analysis in Communication Research.* New York: Free Press, 1952.

Bligh, M., and G.D. Hess. "The power of leading subtly: Alan Greenspan, rhetorical leadership, and monetary policy." *The Leadership Quarterly* 18 (2007): 87-104.

Brown, L. D., and M.L. Caylor.. A temporal analysis of quarterly earnings thresholds: Propensities and valuation consequences. The Accounting Review 80 (2005), 423-440.

Cho, J.; M.P. Boyle; H. Keum; M.D. Shevy; D.M. Mcleod; D.V. Shah; and Z. Pan. "Media, Terrorism, and Emotionality: Emotional Differences in Media Content and Public Reactions to the September 11th Terrorist Attacks." *Journal of Broadcasting & Electronic Media* 47 (2003): 309-328.

Core, J. E.. "A Review of the Empirical Disclosure Literature: Discussion." *Journal of Accounting and Economics* 31 (2001): 441-456.

Davis, A.K.; J.M. Piger; and L.M. Sedor. "Beyond the Numbers: Managers' Use of Optimistic and Pessimistic Tone in Earnings Press Releases." Working paper available at SSRN: http://ssrn.com/abstract=875399. 2007.

Demers, E., and C.Vega. "Soft Information in Earnings Announcements: News or Noise?" Working paper available at SSRN: http://ssrn.com/abstract=1152326. 2010.

Easton, P. D., and M.E. Zmijewski. "SEC Form 10K/10Q Reports and Annual Reports to Shareholders: Reporting Lags and Squared Market Model Prediction Errors*." Journal of Accounting Research* 31 (1993): 113-129.

Engelberg, J. "Costly Information Processing: Evidence from Earnings Announcements." *Working paper*. Available at SSRN: http://ssrn.com/abstract=1107998. (2008)

Francis, J.; D. Philbrick; and K. Schipper. "Shareholder Litigation and Corporate Disclosures*." Journal of Accounting Research* 32 (1994): 137-164.

Francis, J.; K. Schipper; and L. Vincent. “Expanded Disclosures and the Increased Usefulness of Earnings Announcements.” *The Accounting Review* 77 (2002): 515-546.

Frazier, K. B.; R.W. Ingram; and B.M. Tennyson. “A Methodology for the Analysis of Narrative Accounting Disclosures.” *Journal of Accounting Research* 22(1984): 318-331.

Gordon, E. A.; E. Henry: M. Peytcheva; and L. Sun. “Discretionary Disclosure and the Market Reaction to Restatements.” Working paper, Temple University. 2008.

Hayn, C. “The Information Content of Losses.” *Journal of Accounting and Economics* 20(1995): 125-153.

Henry, E. “Market Reaction to Verbal Components of Earnings Press Releases: Event Study Using a Predictive Algorithm.” *Journal of Emerging Technologies in Accounting* 3(2006): 1-19.

Henry, E. “Are Investors Influenced by How Earnings Press Releases are Written?” *Journal of Business Communicatio*n 45 (2008), 363-407.

Hoskin, R. E.; J.S. Hughes; and W.E. Ricks. “Evidence on the Incremental Information Content of Additional Firm Disclosures Made Concurrently with Earnings.” *Journal of Accounting Research* 24 (1986): 1-32.

Hutton, A. P.; G.S. Miller; and D.J. Skinner. “The Role of Supplementary Statements with Management Earnings Forecasts.” *Journal of Accounting Research* 41 (2003): 867-890.

Kothari, S.P. “Capital Markets Research in Accounting.” *Journal of Accounting and Economics*, 31 (2001): 105-231.

Kothari, S.P.; X. Li; and J. Short. “The Effect of Disclosures by Management, Analysts, and Financial Press on Cost Of Capital, Return Volatility, and Analyst Forecasts: A Study Using Content Analysis.” *The Accounting Review* 84 (2009): 1639-1670.

Lang, M. H., and R.J. Lundholm. “Voluntary Disclosure and Equity Offerings: Reducing Information Asymmetry or Hyping the Stock?” *Contemporary Accounting Research* 17 (2000): 623-662.

Li, F. “The Information Content of Forward-Looking Statements in Corporate Filings-A Naïve Bayesian Machine Learning Approach.” *Journal of Accounting Research* 48 (2010): 1049-1102.

Loughran, T., and B. McDonald. “When is a Liability not a Liability? Textual Analysis, Dictionaries and 10-Ks.” *Journal of Finance*, forthcoming February 2011.

Manning, C. D., and H. Schutze. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press. 1999.

Neuendorf, K. A. *The Content Analysis Guidebook*. Sage Publications. 2002.

Ogilvie, D. M.; P.J. Stone, P. J.; and E.S. Shneidman. “Some Characteristics of Genuine Versus Simulated Suicide Notes," in P. J. Stone; D. C. Dunphy; M. S. Smith; and D. M. Ogilvie (Eds.), *The General Inquirer: A Computer Approach to Content Analysis*. Cambridge: MIT Press. 1966.

Rogers, J.; A.Van Buskirk: and S. Zechman. “Disclosure Tone and Shareholder Litigation.” Working paper, University of Chicago. 2009.

Rutherford, B. “Genre analysis of Corporate Annual Report Narratives: A Corpus Linguistics-Based Approach.” *The Journal of Business Communication* 42 (2005): 349-378.

Sadique, S.; F. In; and M. Veeraraghavan. “Impact of Spin and Tone on Stock Returns and Volatility: Evidence from Firm-Issued Earnings Announcements and the Related Press Coverage.” Working paper, Monash University. 2008.

Tetlock, P. C. “Giving Content to Investor Sentiment: The Role of Media in the Stock Market.” *The Journal of Finance* 62 (2007): 1139–1168.

Tetlock, P.C.: M. Saar-Tsechansky; and S. Macskassy S. “More than Words: Quantifying Language to Measure Firms' Fundamentals.” *The Journal of Finance* 63 (2008): 1437-1467.

U.S. Securities and Exchange Commission. *Final rule: Conditions for use of non-GAAP financial measures* (Releases 33-8176 and 34-47226, File S7-43-02). Available from the U.S. Securities and Exchange Commission Web site, http://www.sec.gov/rules/final/33-8176.htm. 2003.

Vuong, Q. H. “Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses.” *Econometrica* 57 (1989): 307–333.

Xuejun Li, E., and K. Ramesh. “Market Reaction Surrounding the Filing of Periodic SEC Reports.” *The Accounting Review* 84(2009): 1171-1208.

TABLE 1
List of Variables

CAR	cumulative abnormal returns from day *t*-1 to day *t*+1 around earnings announcement date
UE	unexpected earnings per share (Compustat variable *EPSPXQ*) in quarter *t* minus earnings per share in quarter *t*-4, scaled by price in *t*-4
ASSETS	total assets in period *t* (Compustat variable *AT*), in millions of dollars.
SIZE	Log of *ASSETS*
NUM_WORDS	count of total words in the disclosure
FD_SCORE	*(POSITIVE – NEGATIVE)/ (POSITIVE + NEGATIVE) where POSITIVE and NEGATIVE* refer to the word count frequency based on the positive and negative words in the FD word list, respectively
DC_SCORE	*(POSITIVE – NEGATIVE)/ (POSITIVE + NEGATIVE) where POSITIVE and NEGATIVE* refer to the word count frequency based on the positive and negative words in the Diction word list, respectively
GI_SCORE	*(POSITIVE – NEGATIVE)/ (POSITIVE + NEGATIVE) where POSITIVE and NEGATIVE* refer to the word count frequency based on the positive and negative words on the General Inquirer word list, respectively
LM_SCORE	*(POSITIVE – NEGATIVE)/ (POSITIVE + NEGATIVE) where POSITIVE and NEGATIVE* refer to the word count frequency based on the positive and negative words on the Loughran-McDonald word list, respectively
NEG_FD	frequency count of negative words on the FD wordlist scaled by *NUM_WORDS*
POS_FD	frequency count of positive words on the FD wordlist scaled by *NUM_WORDS*
NEG_DC	frequency count of negative words on the Diction wordlist scaled by *NUM_WORDS*
POS_DC	frequency count of positive words on the Diction wordlist scaled by *NUM_WORDS*
NEG_GI	frequency count of negative words on the General Inquirer wordlist scaled by *NUM_WORDS*
POS_GI	frequency count of positive words on the General Inquirer wordlist scaled by *NUM_WORDS*
NEG_LM	frequency count of negative words on the Loughran-McDonald wordlist scaled by *NUM_WORDS*
POS_LM	frequency count of positive words on the Loughran-McDonald wordlist scaled by *NUM_WORDS*
LOSS	indicator = 1 if earnings < 0

All continuous variables are winsorized at 1 and 99 percentiles.

TABLE 2
Sample
(*N*= 29,712)

Panel A. Descriptive Statistics
Variables are defined in Table 1.

Variable	Mean	Median	Std Dev	Maximum	Minimum
CAR	-0.001	-0.002	0.074	0.231	-0.232
UE	0.004	0.002	0.038	0.201	-0.158
LOSS	23.68	N/A	N/A	N/A	N/A
ASSETS ($Millions)	4,299	607	13,297	103,377	9
NUM_WORDS	2,019	1,765	1,134	6,645	398
FD_SCORE	0.429	0.462	0.261	1.000	-1.000
DC_SCORE	0.044	0.043	0.346	1.000	-1.000
GI_SCORE	0.334	0.337	0.153	1.000	-0.302
LM_SCORE	-0.037	-0.067	0.348	1.000	-1.000
NEG_FD	0.008	0.007	0.004	0.020	0.001
NEG_DC	0.012	0.010	0.006	0.030	0.002
NEG_GI	0.037	0.036	0.010	0.067	0.016
NEG_LM	0.014	0.012	0.008	0.038	0.001
POS_FD	0.021	0.020	0.009	0.046	0.005
POS_DC	0.012	0.011	0.005	0.030	0.003
POS_GI	0.074	0.073	0.014	0.116	0.047
POS_LM	0.012	0.011	0.005	0.026	0.003

Panel B- Frequency of Observations By Year

File date year	Frequency	Percent
2004	3,515	11.83
2005	13,027	43.84
2006	13,170	44.33
	29,712	100.00

Sample selection period begins in August 2004, when the coding of 8-K filings as an earning press release took effect. Beginning in 2003, the SEC requires that earnings press releases – which are not a mandatory disclosure – must, if issued, be filed with SEC on Form 8-K (SEC 2003).

TABLE 3
Correlation Matrix – Pearson Correlation Coefficients
(*N*= 29,712)

	UE	*SIZE*	*NUM_WORDS*	*FD_SCORE*	*DC_SCORE*	*GI_SCORE*	*LM_SCORE*	*NEG_FD*	*NEG_DC*	*NEG_GI*	*NEG_LM*	*POS_FD*	*POS_DC*	*POS_GI*	*POS_LM*
CAR	0.122 ***	0.054 ***	-0.011	0.109 ***	0.072 ***	0.046 ***	0.092 ***	-0.079 ***	-0.063 ***	-0.041 ***	-0.068 ***	0.088 ***	0.041 ***	0.022 ***	0.058 ***
UE		-0.007	-0.008	0.081 ***	0.012 *	0.009	0.051 ***	-0.056 ***	-0.001	-0.012 *	-0.020 ***	0.064 ***	0.014 *	-0.003	0.061 ***
SIZE			0.524 ***	0.055 ***	0.177 ***	-0.088 ***	-0.024 ***	0.008	-0.200 ***	0.129 ***	0.015 **	0.090 ***	0.065 ***	0.056 ***	-0.022 ***
NUM_WORDS				0.005	0.017 **	-0.198 ***	-0.130 ***	-0.014 *	-0.019 ***	0.187 ***	0.110 ***	-0.034 ***	0.020 ***	-0.071 ***	-0.052 ***
FD_SCORE					0.399 ***	0.186 ***	0.437 ***	-0.689 ***	-0.294 ***	-0.238 ***	-0.295 ***	0.633 ***	0.293 ***	-0.026 ***	0.323 ***
DC_SCORE						0.341 ***	0.604 ***	-0.114 ***	-0.762 ***	-0.341 ***	-0.531 ***	0.498 ***	0.676 ***	0.121 ***	0.309 ***
GI_SCORE							0.377 ***	-0.214 ***	-0.314 ***	-0.843 ***	-0.431 ***	0.084 ***	0.242 ***	0.571 ***	0.105 ***
LM_SCORE								-0.186 ***	-0.548 ***	-0.407 ***	-0.753 ***	0.474 ***	0.331 ***	0.077 ***	0.535 ***
NEG_FD									0.099 ***	0.203 ***	0.160 ***	0.000	-0.054 ***	-0.078 ***	-0.059 ***
NEG_DC										0.415 ***	0.704 ***	-0.335 ***	-0.150 ***	0.026 ***	-0.031 ***
NEG_GI											0.542 ***	-0.172 ***	-0.127 ***	-0.069 ***	-0.006
NEG_LM												-0.285 ***	-0.110 ***	0.005	0.010
POS_FD													0.426 ***	-0.110 ***	0.447 ***
POS_DC														0.265 ***	0.460 ***
POS_GI															0.173 ***

Variables are defined in Table 1.

TABLE 4
Regression of Cumulative Abnormal Returns on Unexpected Earnings and Tone Scores and
Results of Vuong's Test of Relative Information Content
(*N*= 29,712)

$CAR = \alpha + \beta_1 UE + \beta_2 SIZE + \beta_3 X\_SCORE + \varepsilon$
Where all variables are defined in Table 1, and *X_SCORE* refers to the four alternative tone scores.

Panel A. Regression of Abnormal Returns (*CAR*) on Unexpected Earnings and Tone Scores

	Coefficient (t-Stat)	**Coefficient (t-Stat)**	**Coefficient (t-Stat)**	**Coefficient (t-Stat)**
Intercept	-0.023*** (-11.022)	-0.011*** (-5.449)	-0.022*** (-9.571)	-0.013*** (-6.352)
UE	0.223*** (15.110)	0.236*** (15.968)	0.237*** (15.968)	0.229*** (15.547)
SIZE	0.002*** (8.300)	0.002*** (7.272)	0.002*** (9.805)	0.002*** (9.624)
FD_SCORE	0.027*** (15.897)			
DC_SCORE		0.013*** (10.479)		
GI_SCORE			0.024*** (8.917)	
LM_SCORE				0.018*** (14.758)
Year Fixed Effects	Yes	Yes	Yes	Yes
Adjusted R2	2.69%	2.20%	2.08%	2.58%

Panel B. Results of Vuong's Test

Competing Models' Tone Measure	Vuong *Z*-statistic	*p*-value
FD_Score vs. DC_Score	5.378	<0.001
FD_Score vs GI_Score	6.235	<0.001
FD_Score vs. LM_Score	1.566	0.117
DC_Score vs GI_Score	1.833	0.067
DC_Score vs. LM_Score	-4.742	<0.001
GI_Score vs LM_Score	-5.546	<0.001

Regressions use robust standard errors, clustered by gvkey. Significance, two-tailed, indicated as follows: *** <0.001, ** < 0.01, * <0.05.

TABLE 5
Regression of Cumulative Abnormal Returns on Unexpected Earnings and Tone Scores, controlling for Loss-making Companies, and Results of Vuong's Test
(*N*= 29,712)

$CAR = \alpha + \beta_1 UE + \beta_2 SIZE + \beta_3 LOSS + \beta_4 X\_SCORE + \varepsilon$
Where all variables are defined in Table 1, and *X_SCORE* refers to the four alternative tone scores.

Panel A. Regression of Abnormal Returns on Unexpected Earnings and Tone Scores, controlling for Loss-making Companies

	Coefficient (t-Stat)	Coefficient (t-Stat)	Coefficient (t-Stat)	Coefficient (t-Stat)
INTERCEPT	-0.011*** (-4.914)	-0.001 (-0.347)	-0.007** (-2.621)	-0.003 (-1.364)
UE	0.187*** (12.709)	0.198*** (13.268)	0.196*** (13.175)	0.197*** (13.317)
SIZE	0.001*** (3.559)	0.001** (3.210)	0.001*** (4.175)	0.001*** (4.891)
LOSS	-0.015*** (-12.197)	-0.015*** (-12.058)	-0.016*** (-13.139)	-0.014*** (-10.687)
FD_SCORE	0.023*** (13.659)			
DC_SCORE		0.007*** (5.025)		
GI_SCORE			0.014*** (5.203)	
LM_SCORE				0.013*** (10.150)
Year Fixed Effects	Yes	Yes	Yes	Yes
Adjusted R^2	3.33%	2.76%	2.76%	3.01%

Panel B. Results of Vuong's Test

Competing Models' Tone Measure	Vuong's *Z*-statistic	*p*-value
FD_Score vs. DC_Score	6.850	<0.001
FD_Score vs GI_Score	6.456	<0.001
FD_Score vs. LM_Score	3.658	<0.001
Diction_Score vs GI_Score	-0.061	0.951
Diction_Score vs. LM_Score	-4.617	<0.001
GI_Score vs LM_Score	-4.188	<0.001

Regressions use robust standard errors, clustered by gvkey. Significance, two-tailed, indicated as follows: *** <0.001, ** < 0.01, * <0.05.

TABLE 6
Comparison of the Tone Scores Decomposed into Positive and Negative Measures
Regressions of Cumulative Abnormal Returns on Unexpected Earnings and Decomposed Tone Scores, controlling for Loss-making Companies
(*N*= 29,712)

$CAR = \alpha + \beta_1 UE + \beta_2 SIZE + \beta_3 LOSS + \beta_4 NEG\_X + \beta_5 POS\_X + \varepsilon$
Where all variables are defined below, and *NEG_X and POS_X refer* to the four alternative tone scores.

Dependent variable: CAR

	FD Coefficient (t-Stat)	**Diction Coefficient (t-Stat)**	**GI Coefficient (t-Stat)**	**LM Coefficient (t-Stat)**
INTERCEPT	0.000	-0.003	0.000	-0.007 **
	(-0.057)	(-1.341)	(-0.083)	(-2.953)
UE	0.185***	0.195***	0.196***	0.194***
	(12.566)	(13.034)	(13.167)	(13.038)
SIZE	0.001***	0.001**	0.001***	0.001***
	(3.565)	(3.240)	(4.078)	(4.562)
LOSS	-0.015***	-0.016***	-0.016***	-0.014***
	(-12.059)	(-12.090)	(-13.107)	(-11.170)
NEG_X	-1.428***	-0.111	-0.169***	-0.350***
	(-12.290)	(-1.274)	(-3.889)	(-5.754)
POS_X	0.485***	0.353***	0.066*	0.776***
	(9.896)	(4.578)	(2.389)	(8.551)
Year Fixed Effects	Yes	Yes	Yes	Yes
Adjusted R2	3.47%	2.75%	2.74%	3.01%

Regressions use robust standard errors, clustered by gvkey. Significance, two-tailed, indicated as follows: *** <0.001, ** < 0.01, * <0.05. **Variables are defined in Table 1**.

TABLE 7
Regression of Cumulative Abnormal Returns on Unexpected Earnings and Tone Scores—unweighted and IDF-weighted -- controlling for Loss-making Companies, and Results of Vuong's Test
(*N*= 29,712)

$CAR = \alpha + \beta_1 UE + \beta_2 SIZE + \beta_3 LOSS + \beta_4 X\_SCORE + \varepsilon$
Where all variables are defined in Table 1, and *X_SCORE* refers to the four alternative tone scores, unweighted and IDF-weighted.

	FD		**DC**		**GI**		**LM**	
	Equal weight	IDF_weight	Equal weight	IDF_weight	Equal weight	IDF_weight	Equal weight	IDF_weight
	Coefficient (t-Stat)	**Coefficient (t-Stat)**	**Coefficient (t-Stat)**	**Coefficient (t-Stat)**	**Coefficient (t-Stat)**	**Coefficient (t-Stat)**	**Coefficient (t-Stat)**	**Coefficient (t-Stat)**
INTERCEPT	-0.011*** (-4.914)	-0.013*** (-5.780)	-0.001 (-0.347)	-0.002 (-1.010)	-0.007** (-2.621)	-0.006** (2.520)	-0.003 (-1.364)	-0.003 (1.270)
UE	0.187*** (12.709)	0.182*** (12.330)	0.198*** (13.268)	0.194*** (13.030)	0.196*** (13.175)	0.193*** (12.960)	0.197*** (13.317)	0.190*** (12.820)
SIZE	0.001*** (3.559)	0.001*** (5.580)	0.001** (3.210)	0.001*** (3.340)	0.001*** (4.175)	0.001*** (4.000)	0.001*** (4.891)	0.001*** (5.480)
LOSS	-0.015*** (-12.197)	-0.016*** (-13.220)	-0.015*** (-12.058)	-0.017*** (-14.080)	-0.016*** (-13.139)	-0.017*** (-14.470)	-0.014*** (-10.687)	-0.017*** (-13.790)
X_SCORE	0.023*** (13.659)	0.023*** (16.330)	0.007*** (5.025)	0.006*** (4.320)	0.014*** (5.203)	0.015*** (5.790)	0.013*** (10.150)	0.014*** (11.250)
Year Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted R2	3.33%	3.59%	2.76%	2.74%	2.76%	2.81%	3.01%	3.11%

TABLE 8
Top Twenty Words Contributing to FD Tone Score More Negative than GI Tone Score

Words included as negative only in the FD wordlist or as positive only in the GI wordlist and occurring with the greatest frequency in 1,485 press releases (top 5 percent of score divergence) with an FD tone score more negative than the GI tone score

Panel A. Words included as negative in FD wordlist but not in GI wordlist and occurring in the most documents among the 1,485 press releases with FD tone score more negative than GI tone score.

	Total documents	Total frequency	IDF Weighting	FD negative	GI negative	DC negative	LM negative
RISKS	1291	3882	0.205	1	0	0	0
UNCERTAINTIES	1288	2731	0.227	1	0	0	0
UNDER	1092	4342	0.395	1	0	0	0
RISK	737	2142	0.938	1	0	1	0
LESS	672	2438	0.891	1	0	0	0
BELOW	398	826	1.277	1	0	0	0
DOWN	344	826	1.398	1	0	0	0
LEAST	231	297	1.969	1	0	0	0
NEGATIVELY	142	202	2.336	1	0	0	1
UNCERTAINTY	132	167	2.531	1	0	0	0
UNCERTAIN	97	102	3.119	1	0	0	0
FALLEN	32	46	4.792	1	0	0	0
SMALLER	31	36	3.455	1	0	0	0
WEAKNESS	30	36	3.335	1	0	1	1
PENALTIES	25	42	3.993	1	0	0	1
LOWEST	24	28	3.895	1	0	0	0
PENALTY	23	33	4.324	1	0	0	1
WEAK	21	23	3.685	1	0	1	1
DOWNTURN	19	20	4.005	1	0	0	1
WEAKNESSES	15	21	4.077	1	0	1	1

Table 8 (Continued)

Panel B. Words included as positive in GI list but not in FD list and occurring with in the most documents among the 1,485 press releases with FD Tone Score More Negative than GI Tone Score

	Total documents	Total frequency	IDF Weighting	FD positive	GI positive	DC positive	LM positive
SHARE	1450	18965	0.101	0	1	0	0
COMPANY	1395	17008	0.050	0	1	0	0
FORWARD	1375	6096	0.074	0	1	0	0
SHARES	1357	7289	0.223	0	1	0	0
INTEREST	1297	13398	0.194	0	1	0	0
EQUITY	1288	7792	0.250	0	1	0	0
BASIC	1276	4186	0.317	0	1	0	0
ACTUAL	1243	1895	0.166	0	1	0	0
COMMON	1231	9511	0.333	0	1	0	0
CONTACT	1175	1352	0.330	0	1	0	0
OUTSTANDING	1152	3882	0.359	0	1	1	1
CONSOLIDATED	1119	4345	0.332	0	1	0	0
COMMISSION	926	1262	0.451	0	1	0	0
VALUE	909	3420	0.578	0	1	0	0
CALL	875	4068	0.476	0	1	0	0
PRIMARILY	838	2163	0.578	0	1	0	0
ACCRUED	738	1442	0.729	0	1	0	0
PAID	693	1314	0.826	0	1	0	0
GAIN	692	3251	0.883	0	1	0	1
ABILITY	679	1442	0.678	0	1	0	0

TABLE 9
Top Twenty Words contributing to FD Tone Score More Positive than GI Tone Score

Words included as positive only in the FD wordlist or as negative only in General Inquirer wordlist and occurring with the greatest frequency in the 1,485 press releases (top 5 percent of score divergence) with an FD tone score more positive than the GI tone score

Panel A. Words included as positive in FD wordlist but not in GI wordlist and occurring in the most documents among the 1,485 press releases with FD tone score more positive than GI tone score

	Total documents	Total frequency	IDF Weighting	FD positive	GI positive	DC positive	LM positive
INCREASE	1342	8560	0.210	1	0	0	0
INCREASED	1308	9879	0.275	1	0	0	0
GROWTH	1240	8488	0.384	1	0	1	0
MORE	1217	4992	0.282	1	0	0	0
STRONG	1081	3549	0.728	1	0	1	1
CERTAIN	1017	3237	0.474	1	0	0	0
UP	958	4553	0.721	1	0	0	0
HIGHER	918	4591	0.776	1	0	0	0
RECORD	775	2827	0.985	1	0	0	0
LEADING	742	1342	0.972	1	0	0	1
INCREASES	741	1875	0.942	1	0	0	0
HIGH	663	1511	1.042	1	0	0	0
MOST	655	1194	0.976	1	0	0	0
ABOVE	565	1277	1.151	1	0	0	0
GREW	491	1425	1.634	1	0	0	0
IMPROVEMENTS	446	725	1.664	1	0	0	1
LARGEST	436	610	1.473	1	0	0	0
SOLID	430	2271	1.661	1	0	0	0
INCREASING	409	610	1.512	1	0	0	0
OPPORTUNITIES	402	566	1.444	1	0	0	1

Table 9 (Continued)

Panel B. Words included as negative in GI wordlist but not in FD wordlist and occurring in the most documents among the 1,485 press releases with FD tone score more positive than GI tone score.

	Total documents	Total frequency	IDF Weighting	FD negative	GI negative	DC negative	LM negative
EXPENSE	1395	11141	0.219	0	1	0	0
TAXES	1364	8020	0.327	0	1	0	0
COST	1306	6398	0.305	0	1	0	0
COSTS	1274	10484	0.338	0	1	0	0
TAX	1256	11399	0.393	0	1	0	0
PRESS	1201	2639	0.247	0	1	0	0
LOSS	1162	15209	0.335	0	1	1	1
DIFFER	1156	1613	0.233	0	1	0	0
CAPITAL	1143	5080	0.383	0	1	0	0
DEPRECIATION	1077	4082	0.742	0	1	0	0
SERVICES	989	8678	0.611	0	1	0	0
EXCLUDING	877	5215	0.931	0	1	0	0
SERVICE	815	4054	0.738	0	1	0	0
CHARGES	769	6180	0.953	0	1	0	0
SHORT	660	1344	1.056	0	1	0	0
LIMITED	583	901	0.754	0	1	0	0
CHARGE	580	2436	1.083	0	1	0	0
FOREIGN	562	2124	1.534	0	1	0	0
COMPETITIVE	542	723	1.063	0	1	0	0
INVOLVE	487	523	0.946	0	1	0	0

Figure 1
Comparison of the Power of the Tone Scores Used in Regressions of Cumulative Abnormal Returns on Unexpected Earnings and Tone Measures for 250 Randomly-Selected Samples of Varying Sizes

For randomly selected samples of 50, 100, 250, 500, 1000, and 2000 the following equation was estimated: $CAR = \alpha + \beta_1 UE + \beta_2 SIZE + \beta_3 LOSS + \beta_4 SCORE_{tone} + \varepsilon$ Where all variables are defined in Table 1, and $SCORE_{tone}$ refers to the four alternative tone scores.

Figure 1A - Rejection rates (β_4) for unweighted tone scores

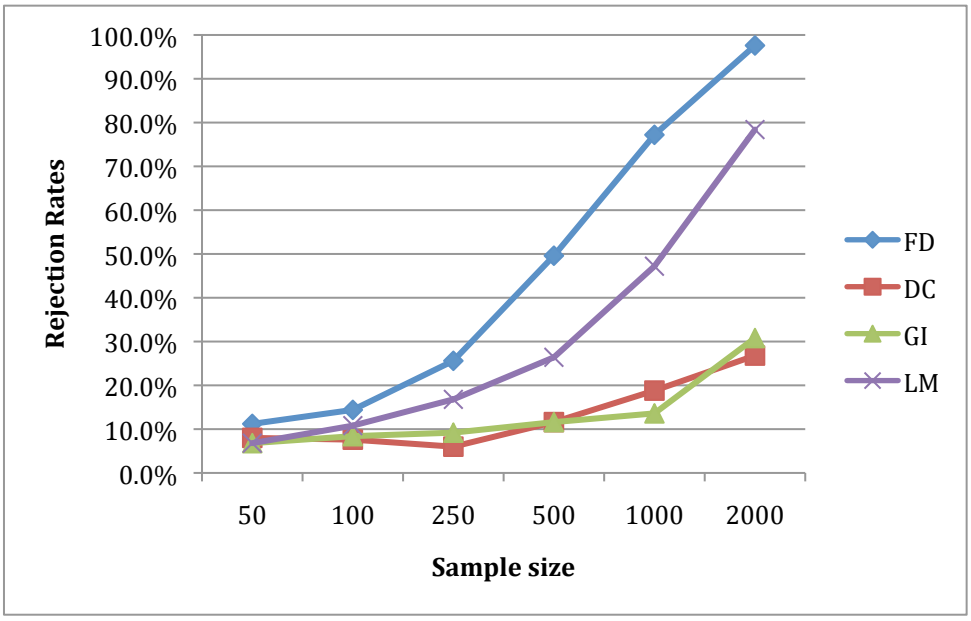
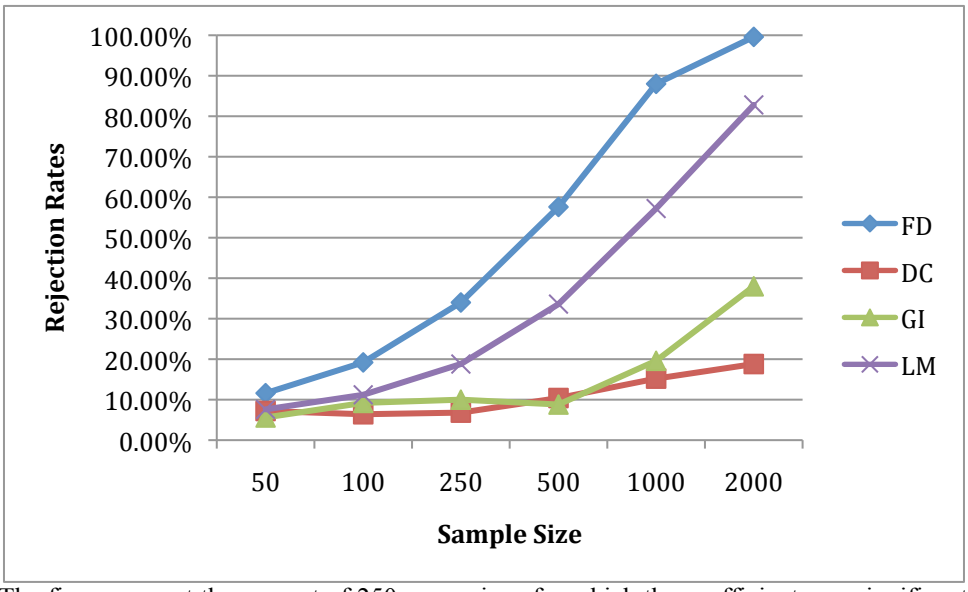


Figure - 1B - Rejection rates (β_4) for IDF-weighted tone scores



The figures report the percent of 250 regressions for which the coefficient was significant at or below the 5 percent level. **Variables are defined in Table 1**. For the IDF-weighted measures, a separate IDF weighting is calculated for each of the 1,500 different subsamples.

Figure 2
Comparison of the Power of the Tone Scores Used in Regressions of Cumulative Abnormal Returns on Unexpected Earnings and Tone Measures for 250 Randomly-Selected Samples of Varying Sizes

For randomly selected samples of 50, 100, 250, 500, 1000, and 2000 the following equation was estimated: $CAR = \alpha + \beta_1 UE + \beta_2 SIZE + \beta_3 LOSS + \beta_4 NEG\_X + \beta_4 POS\_X + \varepsilon$
Where all variables are defined in Table 1, and *NEG_X and POS_X* refer to the four alternative negative and positive tone scores (FD,GI, DC, and LM.

Figure 2A – Rejection rates for NEG_X (β_4) for various sample sizes.

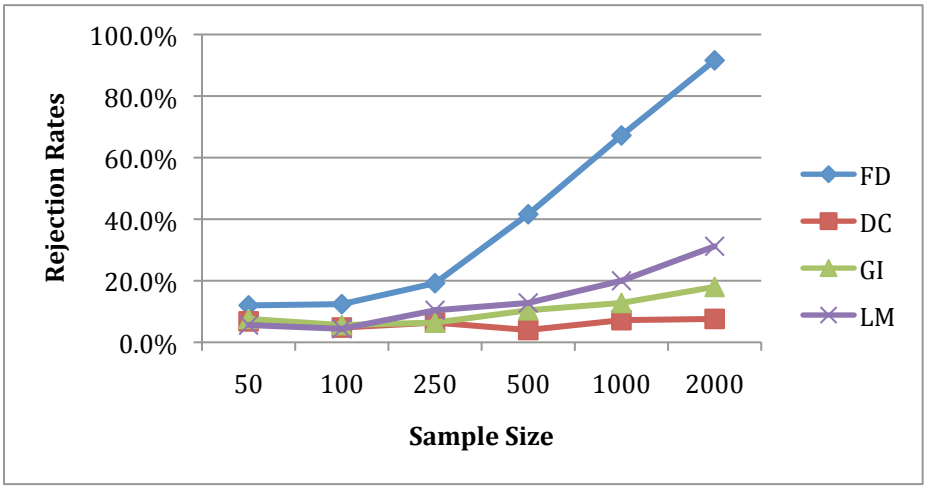
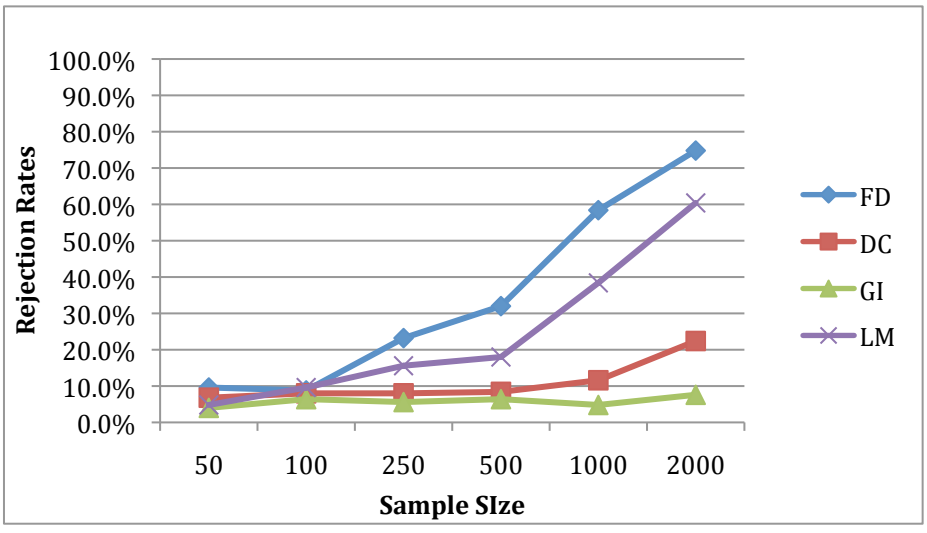


Figure 2B - Rejection rates for POS_X (β_5) for various sample sizes



The figures report the percent of 250 regressions for which the coefficient was significant at or below the 5 percent level. **Variables are defined in Table 1**.

Appendix A. Document-weightings

This Appendix presents two examples of document-weighting, specifically the *idf* (inverse document frequency), in developing measures of negative tone. Two of the basic tasks in computational linguistics are content analysis and information retrieval, i.e. search. To find or retrieve a document of interest from among a corpus,[29] it is necessary to have some measure of the content of each document in the corpus, which is then combined with a weighting that identifies the document most likely to be of interest for the search. Measures of content are typically based on proportional word frequency. Weightings used to improve the precision of a search by ranking documents within a corpus are typically based on frequency of occurrence across all documents in a corpus; less frequently occurring words can better discriminate among documents that are most closely related to a specific search term.

Idf weightings are designed to favor words that occur in fewer documents because those words improve the ability to rank documents according to their relevance to a particular set of search words and thus to improve the precision of a search algorithm. The *idf* weight is defined as $\log(N/n_j)$ where N is the total number of documents and n_j is the number of documents that contain the search word. A search word that appears in many documents would contribute little to ranking documents' relevancy to the search, and a search word that appeared in every document in the corpus would contribute nothing; its *idf* weight would be $\log(1)=0$.

Loughran and McDonald (2009) argue that *idf* weightings should be used in measures of qualitative information to mitigate the problem of words that are mistakenly classified as negative in the GI wordlist. We argue that the problem is more directly solved by removing the improperly classified words from the list.

Idf weightings do not solve the problem of misclassified words in a tone wordlist. *Idf* weightings dampen the impact of misclassified words in a tone wordlist only when the words are

[29] In computational linguistics, the term "corpus" is used to refer to the group or sample of documents being analyzed.

common within a corpus but magnify the impact of misclassified words in a tone wordlist when the words are rare within the corpus. *Idf* weightings obscure the relative proportion of a document that pertains to negative or positive information, and *idf* weightings are dependent on sample selection.

The *wf-idf* weight combines a word's frequency count within each document and a measure of its frequency across all documents in a corpus (its inverse document frequency) into a single weight. For each word in a document, the *wf-idf* weight is the product of its word frequency *wf* weight times the *idf* weight. When the word frequency weight is calculated as the log of the word frequency count, the *wf-idf* measure can be formed as shown below (from Manning and Schutze, 2002, 543).

$$wf\text{-}idf = \begin{cases} (1 + \log(wf_{i,j}))\log(N/n_j) & \text{if } wf_{i,j} >= 1 \\ 0 & \text{if } wf_{i,j} = 0 \end{cases} \qquad (A1)$$

The first example in this appendix illustrates how *idf* weights make tone measures extremely sensitive to the sample corpus. The example uses a hypothetical set of three documents, each with a total of 100 words, with the word-frequency count matrix shown in Panel A of Table A1. In a word-frequency count matrix, documents are represented as vectors of word frequency counts. For example, the word "less" appears 10 times in Document 1, and the word "lower" appears 10 times in Document 2 and 1 time in Document 3. The words "less" and "lower" are the only two negative words in the corpus. The remaining words in the document would each be represented as a separate column of the matrix, but for presentation purposes, we show them here in a single column labeled "another." The *idf* weight for each word is presented at the bottom of the column. The *idf* weight for the word "less" is 1.10 because it appears in only one of the three documents, (*idf* = ln(3/1) = 1.10. The *idf* weight for the word "lower" is only 0.41, calculated as (*idf* = ln(3/2)), because it appears in two of the three documents.

Two negative tone measures for each of the three documents are shown in Panel B of Table A1, the first measure using equal weighting and the second measure using the combined *wf-idf*

weighting according to Equation A1 above.[30] Because D1 and D2 have the identical proportion of negative words, the equal-weighted measure of negative tone is identical for the two documents. When the negative tone measure adds an *idf* weighting, D1 is scored as 2.7 times more negative than D2, but the increase in D2's negative tone measure is solely because the word "lower" appears once in D3. If the corpus had excluded D3 or if D3 had not contained the word "lower", then the *wf-idf* negative tone measure for D2 would have been identical to that of D1.

Table A1. Sensitivity of IDF-weighted tone measures to composition of sample corpus

Panel A. Word-frequency count matrix for set of three hypothetical documents

	"LESS"	*"LOWER"*	*"Another"*	Total words
D1	10	0	90	100
D2	0	10	90	100
D3	0	1	99	100
idf weight (N/ n_j)	1.10	0.41	0.00	

Panel B. Negative tone measures, using equal weighting versus *idf* -weighted measures

	Negative tone, measured as:	
	Equal-weighted *wf*. Negative words as a proportion of total words.	Inverse-document-frequency-weighted *wf-idf* (Equation A1).
D1	0.100	3.628
D2	0.100	1.339
D3	0.010	0.405

The next example illustrates how the *idf* weights serve to correct for errors in mistakenly classified words that commonly occur across a corpus. Table A2, Panel A shows the word-frequency count matrix for a hypothetical three-document corpus. Two alternative wordlists are used to create negative tone measures for the documents. The first wordlist includes only two negative words "less" and "lower," while the second wordlist incorrectly classifies the word "division" as negative. Based on the equal-weighted tone measures shown in Panel B, the incorrect wordlist scores the documents as far more negative than the correct wordlist (four times more negative than the correct wordlist for

[30] Loughran and McDonald (2009) additionally scale the word frequency by the total words in the document to control for differences in length. For simplicity, our examples use documents of identical lengths.

D1 and D2, and 31 times more negative for D3). An *idf* weighting serves to negate the effect of the incorrectly classified word because it is very common in the corpus. Specifically, because the incorrectly classified word "division" appears in all three documents, its *idf* weight is zero (*idf* = log (3/3)), so the *wf-idf* measure eliminates the word from the negative tone measure. Using the *wf-idf* measure, therefore, the incorrect wordlist can produce exactly the same tone measure as the correct wordlist. Of course, correcting the list by removing the improperly classified word "division" from the negative word list would be a simpler and more direct solution.

Table A2. *Idf* weights correct tone measures for errors in classifying words that appear frequently in a corpus

Panel A. Word-frequency count matrix for set of three hypothetical documents

Included in correct negative list	Yes	Yes	No	No	
Included in incorrect negative list	Yes	Yes	Yes	No	
	"LESS"	*"LOWER"*	*"DIVISION"*	*Another*	Total
D1	10	0	30	60	100
D2	0	10	30	60	100
D3	0	1	30	69	100
idf weight (N/ n_j)	1.10	0.41	0.00	0.00	

Panel B. Negative tone measures, using equal-weighted and *idf* -weighted measures applied to the correct wordlist versus the incorrect wordlist

	Negative tone, measured as:					
	Equal-weighted *wf*. Negative words as a proportion of total words.			Inverse-document-frequency-weighted *wf-idf* (Equation A1).		
	Measure using correct list	Measure using incorrect list	**Ratio of Incorrect to Correct**	Measure using correct list	Measure using incorrect list	**Ratio of Incorrect to Correct**
D1	0.100	0.400	**4**	3.628	3.628	**1**
D2	0.100	0.400	**4**	1.339	1.339	**1**
D3	0.010	0.310	**31**	0.405	0.405	**1**

In summary, although *idf* weights can correct tone measures for misclassified words that appear frequently in a corpus, a far more direct solution would be simply to remove the mistakenly classified words from the word list. In addition, *idf* weights do not fully solve the problem of

misclassification. First, adding *idf* weights obviously does not serve to include negative words that were mistakenly omitted from a wordlist. Second, when a misclassified words occurs infrequently across a corpus, *idf* weights exacerbate their impact on the negative tone measure. Thus, use of *idf* weights potentially offers only limited benefits but comes with a significant cost: in addition to being more difficult to implement, it reduces the transparency and replicability of studies using word-count based measures of tone.

Appendix B

Excerpt From a Press Release Where FD Tone Score Much More Positive Than GI Tone Score

Words appearing as positive on the FD wordlist only are marked with italics, shading, and double underline. Words appearing as negative on the GI wordlist only are marked with bold, single underline (except for the word "Monster," which appears as negative only on the GI wordlist). Words appearing as on both FD and GI positive lists or on both FD and GI negative list are not identified

---

Monster Worldwide Reports Fourth Quarter and Full Year 2004 Results

2004 Fourth Quarter Financial Highlights

- Diluted EPS *Increases* 82% to $0.20 on 45% Year over Year Revenue Gain
- Monster **Division**'s Revenue Reaches $172 Million, *Up* 64% From a Year Ago
- Deferred Revenue at Monster **Division** Sets New *Record* of $230 Million
- Gross Cash Position *Increases* 69% Sequentially to $198 Million
- Monster **Division**'s Operating Margin *Expands* to 23%, *Up* From 18.2% in Q3 2004

New York, February 1, 2005 – Monster Worldwide, Inc. (NASDAQ: MNST), the parent company of the *leading* global online careers property, Monster®, the world's *largest* Yellow Pages advertising agency, and one of the world's *largest* Recruitment Advertising agency networks, today reported financial results for the fourth quarter ended December 31, 2004.

Fourth Quarter Results

Monster Worldwide's total revenue *increased* 45% to $236.8 million in the fourth quarter of 2004 from $163.2 million in the comparable quarter last year. The *strong* revenue *growth* was driven by exceptional global sales performance at the Monster **division**, a *solid* *increase* in the Company's Advertising & Communications business in North America and contributions from acquisitions made earlier in the year. The Monster **division** recorded revenue of $172.2 million, a 64% *increase* over last year's fourth quarter level of $104.9 million. Sequentially, the **division**'s revenues grew 9% over the $157.7 million reported in the third quarter of this year. Organic revenue *growth* over the 2003 fourth quarter was 28%. The Monster **division**'s deferred revenue balance was a *record* $230.1 million, an *increase* over the $195.4 million for this year's third quarter and a 50.2% gain over the $153.2 million recorded in last year's fourth quarter. Consolidated net income for the fourth quarter *doubled* to $24.5 million from the $12.1 million reported in the fourth quarter of 2003. Consolidated net income includes a gain on the sale of the Company's US Motivation business, significantly offset by a **loss** on the sale of *certain* continental European AdComms operations, both of which are reported as discontinued operations. Diluted earnings per share were $0.20 for the current quarter versus $0.11 for last year's comparable quarter.

"Our *strong* global performance in the fourth quarter capped off a terrific 2004 for Monster Worldwide," said Andrew McKelvey, Chairman and Chief Executive Officer of Monster Worldwide. "We executed on our aggressive plan to *increase* profits and *expand* the Monster franchise. As these results demonstrate, we *successfully* improved our operational efficiencies, while *growing* our client base, and enhancing our breadth of offerings."

"By going after small and medium-sized businesses and taking market share from newspapers and by *expanding* our global footprint in key markets such as Europe and India, we unleashed the true potential of Monster in 2004. With our world-class sales operation, innovative products and services, including our successful eCommerce channel, and the continued migration of recruitment to the Internet throughout the world, we are well positioned to continue our *solid* *growth* as we enter 2005."

Cash generated from operating activities was $44.1 million in the fourth quarter of 2004, a $10.7 million *increase* over the $33.4 million of cash from operating activities in the comparable 2003 quarter. The Company's cash position was $198.1 million at December 31, 2004. Cash and cash equivalents at September 2004 and December 2003 were $117.2 million and $142.3 million, respectively. Net cash at December 31, 2004 *increased* to $134.9 million from the $60.2 million reported at September 30, 2004.