Stanford GSB Research Paper No. 2060
Rock Center for Corporate Governance Working Paper No. 83

Detecting Deceptive Discussions in Conference Calls

David F. Larcker
Anastasia A. Zakolyukina

July 2010

STANFORD
GRADUATE SCHOOL OF BUSINESS

The Rock Center
for Corporate Governance
STANFORD UNIVERSITY

A joint initiative of Stanford Law School and the Graduate School of Business

Detecting Deceptive Discussions in Conference Calls*

David F. Larcker† Anastasia A. Zakolyukina‡

July 29, 2010

Abstract

We estimate classification models of deceptive discussions during quarterly earnings conference calls. Using data on subsequent financial restatements (and a set of criteria to identify especially serious accounting problems), we label the Question and Answer section of each call as "truthful" or "deceptive". Our models are developed with the word categories that have been shown by previous psychological and linguistic research to be related to deception. Using conservative statistical tests, we find that the out-of-sample performance of the models that are based on CEO or CFO narratives is significantly better than random by 4%- 6% (with 50% - 65% accuracy) and provides a significant improvement to a model based on discretionary accruals and traditional controls. We find that answers of deceptive executives have more references to general knowledge, fewer non-extreme positive emotions, and fewer references to shareholders value and value creation. In addition, deceptive CEOs use significantly fewer self-references, more third person plural and impersonal pronouns, more extreme positive emotions, fewer extreme negative emotions, and fewer certainty and hesitation words.

*We would like to thank Thomas Quinn for his help in securing the FactSet data and Daniel Jurafsky, Jerome Friedman, Maria Correia, Maria Ogneva, Miguel Angel Minutti Meza, and participants at the Transatlantic Doctoral Conference 2010 at the London Business School for helpful discussions.

†Graduate School of Business, Rock Center for Corporate Governance, Stanford University; email:Larcker_David@gsb.stanford.edu

‡Graduate School of Business, Stanford University; email: aaz@stanford.edu

1. Introduction

Assessing whether reported financial statements are intentionally misstated (or manipulated) is of considerable interest to researchers, creditors, equity investors, and governmental regulators. Prior research has used a variety of accounting-based models to uncover manipulations (e.g., Jones [1991], Dechow and Dichev [2002], McNichols [2000], Dechow et al. [2010]). In addition, professional organizations, such as Audit Integrity Inc., have developed heuristics based on accounting relations to provide warning signs of manipulation.[1] Despite extensive prior research, the ability of these models to identify and predict accounting manipulations is modest.

In this paper, we take a different approach to the prediction of financial statement manipulations by analyzing linguistic features present in answers of CEOs and CFOs during quarterly earnings conference calls. In particular, we examine the Question and Answer (Q&A) narrative in conference calls for linguistic features that predict "deceptive" reporting of financial statements. Our study is grounded in psychology and deception detection research, which finds that the language composition of true narratives differs from that of false narratives. Our primary assumption is that CEOs and CFOs know whether financial statements have been manipulated, and their spontaneous and (hopefully) unrehearsed narratives provide cues that can be used to identify lying or deceitful behavior.[2]

Using the electronic transcripts of quarterly conference calls from FactSet Research Systems Inc. and restatements identified by Glass, Lewis and Co., we build prediction models for the likelihood of deception in quarterly financial statements. We label conference call narratives as "deceptive" if they involve substantial subsequent restatement of net income

[1] See: http://www.auditintegrity.com/.

[2] Our approach does not use the formal presentation text from conference calls because this part of the presentation has been rehearsed by executives, and reviewed by the general counsel, the outside legal counsel, and the investor relations function. As discussed later, we believe that this formal text is an inferior corpus for detecting deception relative to the more spontaneous Q&A discussion.

and are associated with more severe types of restatements such as the disclosure of a material weakness, the change of an auditor, a late filing, or a From 8-K filing. In out-of-sample tests, we find that our linguistic classification models based on CFO (CEO) narratives perform significantly better than a random classifier by 4% - 6% with the 50% - 65% of narratives correctly classified. We also find that the model based on linguistic (word) categories has significantly better predictive performance compared to a model based on discretionary accruals.

In terms of linguistic features of deceptive narratives, we find that deceptive CEOs and CFOs use more references to general knowledge, fewer non-extreme positive emotions words, fewer references to shareholders value and value creation. However, we also find substantial differences between CEOs and CFOs. Deceptive CEOs use significantly fewer self-references, more third person plural and impersonal pronouns, fewer extreme negative emotions words, more extreme positive emotions words, fewer certainty words, and fewer hesitations. In contrast, deceptive CFOs do not have extreme negative emotions and extreme positive emotions words significantly associated with deception. These results are generally consistent with prior theoretical and empirical studies of deception in psychology and linguistics.

Overall, our results suggest that linguistic features of CEOs and CFOs in conference call narratives can be used to identify deceptive financial reporting. Unlike extant accounting-based models that impose stringent data requirements, this linguistic approach can be applied to any company that has a conference call. It is also useful to highlight that predicting accounting manipulation is an extremely difficult task, and high levels of performance are unlikely for this initial study (i.e., the proverbial "needle in the haystack" problem). Despite this caveat, we believe that our initial results are sufficiently interesting that it is worthwhile for researchers to consider linguistic features when attempting to measure the quality of reported financial statements.

The reminder of the paper consists of seven sections. Section 2 provides a review of

how the prior accounting and finance work measures linguistic features of press releases, formal Securities and Exchange Commission (SEC) filings, and other similar text documents. Section 3 discusses the theoretical justification for the word categories that we use to estimate our classification models. The sample construction is discussed in Section 4, and measurement and econometric choices are developed in Section 5. The primary results for our linguistic prediction models are presented in Section 6, and extensions in Section 7. Concluding remarks, limitations, and suggestions for future research are provided in Section 8.

2. Prior Research Analyzing Linguistic Features

Recent papers in accounting and finance analyze various linguistic features inherent in formal corporate disclosures (e.g., Demers and Vega [2010], Li [2006], Li [2008], Li [2010], Loughran et al. [2009]), press releases (e.g., Davis et al. [2007], Henry and Leone [2009]), media news (e.g., Tetlock [2007], Tetlock et al. [2008], Core et al. [2008]), and internet message boards (e.g., Antweiler and Frank [2004], Das and Chen [2007]). These studies essentially measure the positive (optimistic) or negative (pessimistic) tone.[3] However, as we discuss below, they differ in the linguistic cues under consideration and techniques for extracting them. For example, some studies count presence of particular words, whereas others analyze overall tone of the message. Researchers use hand-collected lists of words, simple word counts from psychosocial dictionaries, and estimates produced by natural-language processing classifiers.

Some prior work assumes that a carefully selected list of words can capture a particular linguistic characteristic. For example, Li [2006] examines risk sentiment of annual 10-K filings; where risk sentiment is measured by counting words related to risk ("risk", "risks",

[3]Studies in accounting and finance typically analyze the formal text of the 10-K, MD&A sections, press releases, etc. This is potentially problematic because these texts are highly structured and likely reviewed by the legal and investor relations staff. As a result, they may not be diagnostic with respect to deception, which is the focus of our research.

"risky") and uncertainty ("uncertain", "uncertainty", "uncertainties"). Core et al. [2008] analyze newspaper articles about CEO compensation and identify articles that have negative tone by keywords. Similarly, Loughran et al. [2009] suggest a list of ethics-related terms that they search for in 10-K annual reports.

Although a useful approach, hand-collected lists of words can be confounded by potential subjectivity and miss important dimensions that are captured by more comprehensive psychosocial dictionaries and automatic classifiers. However, an advantage of this approach is that hand collection forces a researcher to identify the linguistic dimension of interest and the precise words that are related to this construct.

Another strand of this literature employs psychosocial dictionaries to count words that reflect particular characteristics of the text such as General Inquirer (GI) or Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al. [2007]). For instance, Tetlock [2007] examines investor sentiment extracted from the "Abreast of the Market" column in the Wall Street Journal by measuring the pessimism index that is composed of mostly negative and weak words from the GI dictionary. Kothari et al. [2008] also use GI to count negative and positive words in disclosures by management, analysts' reports and business news. Davis et al. [2007] measure linguistic style (tone) in earnings press releases using software package DICTION.[4] They count the fraction of words that are optimistic (praise, satisfaction, and inspiration) and pessimistic (blame, hardship, denial).

Similarly using LIWC word categories as our study, Li [2008] examines disclosures made in annual reports counting linguistic features related to obfuscation such as relative frequency of self-reference words, causation words, positive emotions words, and future tense verbs. He finds that more causation words, less positive words, and more future tense verbs are associated with less persistent positive earnings, which is consistent with an obfuscation (probably lying) story.

[4]See: http://www.dictionsoftware.com/index.php.

There are several issues with using psychosocial dictionaries. Perhaps most important, word counting programs do not differentiate between several meanings of words with the same appearance. Pure word counting also does not categorize combinations of words (or phrases) that might possess different meanings from the constituent words. Equally problematic, most of the general dictionaries are not compiled for analyzing business communication. However, assuming that word counting is valid for the research setting, this approach is parsimonious, replicable, and transparent.

Another approach is to apply text classifiers from computational linguistics such as Naive Bayesian algorithm. For example, Antweiler and Frank [2004] examine 1.5 million messages posted on Yahoo!Finance and Raging Bull for 45 companies in the Dow Jones Industrial Average and the Dow Jones Internet Index. Messages are automatically classified into {BUY, HOLD, SELL} category. Similarly, Balakrishnan et al. [2010] use text classification to assign manufacturing firms as out-/under-performing based on narrative disclosures in their 10-K filings. With respect to association between the tone of forward-looking statements in the MD&A of 10-K and 10-Q filings, Li [2010] concludes that the tone measure estimated by Naive Bayesian classifier is significantly positively associated with future performance, whereas the tone measures extracted using traditional dictionaries (Diction, GI, LIWC) are not associated with future performance.

Despite its sophistication, one of the issues with automatic classification is that the classifiers and dictionaries produced may be highly sample specific, and thus have weak diagnostic power on new data. This deficiency can be somewhat alleviated by incorporating domain-specific knowledge. For example after automatically generating the word list, a researcher can complete the list with domain-specific synonyms from a lexical database such as WordNet.[5]

Prior accounting and finance research has provided a number of interesting correlations

[5]See: http://wordnet.princeton.edu/.

between linguistic cues and firm performance outcomes. However, with the possible exception of obfuscation analysis by Li [2008], there is little prior work on using linguistic features to identify deceptive reporting behavior by corporate executives. The purpose of this paper is to use linguistic analysis to develop a predictive model for deception or lying about financial performance by CEOs and CFOs.

3. Development of Word Categories

3.1 Theoretical background

We base our theoretical development of word categories on the extensive review and analysis provided by Vrij [2008].[6] As discussed in Vrij [2008], the common theoretical perspectives used to explain an individual's nonverbal behavior during deception may be also used to explain verbal content of a deceptive speech. These four perspectives include emotions, cognitive effort, attempted control, and lack of embracement theories.

First, the emotions perspective hypothesizes that deceivers feel guilty and are afraid to be caught in a deceptive act. Consequentially, they can experience negative emotions that are manifested in both negative comments and negative affect. Deceivers are also likely to use general terms and not to refer explicitly to themselves. As a result of this dissociation, their statements are often short, indirect, and evasive.

Second, proponents of the cognitive effort perspective argue that fabricating a lie is difficult. That is, if a liar has little or no opportunity to prepare or rehearse, his/her verbal statements are likely to lack specific detail and include more general terms. Similar to the emotions perspective, this cognitive perspective implies fewer self-references and shorter statements. Thus, a liar may sound implausible and non-immediate, telling a story that avoids mentioning any personal experiences.

[6] Vrij [2008] reviews 69 studies (published in English) that examine the verbal behavior of liars and truth-tellers.

Third, control perspective theorists argue that liars avoid producing statements that are self-incriminating. As a result, the content of deceptive statements is controlled so that listeners would not easily perceive it to be a lie. Consistent with the aforementioned theories, this perspective implies general, non-specific language, fewer self-references, short statements with little detail, and more irrelevant information as a substitute for information that the deceiver does not want to provide. For example, a liar speaks with greater caution, using a greater number of unique words to achieve lexical diversity. In contrast, truth-tellers often repeat the information they have provided; such repetition leads to less lexical diversity.

On the other hand, attempted control may also lead to a very smooth speech when a narrative is prepared and rehearsed in advance, whereas truth-tellers often forget (or adapt) what they have said previously.[7] In contrast to the cognitive effort perspective, the attempted control theory implies that well-prepared answers are likely to contain fewer hesitations, more specific statements, and a reduced number of general claims.

Finally, the advocates of the lack of embracement perspective argue that liars appear to lack conviction because they feel uncomfortable when they lie, or they have not personally experienced the supposed claims. This perspective implies that liars use more general terms, fewer self-references, and shorter answers.

Overall, psychological and linguistic theories suggest that liars are more negative and use fewer self-references. However, depending on the theoretical perspective (cognitive effort or attempted control) and whether the questions are expected and the answers are well-rehearsed, the associations between specific linguistic features and deception are ambiguous. The next subsection describes specific verbal cues of deception that we include in our prediction models.

[7]Hence, to gain some insight into conference calls, we discussed this disclosure format with several investor relations consulting firms. They all suggested that a conference call is an important event that involves considerable preparation (and "rehearsal") by the management team on a range of possible questions that are likely to be asked (specifically of the CEO and CFO).

3.2 List of word categories

Although not a specific word category, several papers use response length measured by the number of words as a deception cue (e.g., DePaulo et al. [2003], Newman et al. [2003]). For instance, DePaulo et al. [2003] hypothesizes that liars are less forthcoming than truth-tellers and, as a result, their responses are brief. Similarly, advocates of emotions, cognitive effort, and lack of embracement perspectives argue that deceivers produce shorter statements. In contrast, the attempted control perspective suggests that a falsified story can be well rehearsed, elaborate, and longer. Thus, there is ambiguity about the direction of association between word count and untruthful statements.[8]

Our measurement strategy for the remainder of the word categories is to use well-developed word lists (e.g., LIWC and WordNet) where appropriate. As described below, LIWC is a source for positive and negative emotions words, pronouns, certainty and tentative words , and speech hesitations. We expand some categories by adding synonyms from a lexical database of English WordNet[9] (Table 1, Part B). To establish word categories specific to deception in the conference call setting, we examined ten transcripts for the quarters that have their financial results being subsequently restated. Based on our reading of these transcripts, we create word lists for references to general knowledge, shareholders value, and value creation. The description of word categories, typical words included in each category, prior research supporting the category, and hypothesized signs of association with untruthful narratives are summarized in Table 1.

The literature suggests that the use of first-person singular pronouns implies an individual's ownership of a statement, whereas liars try to dissociate themselves from their words due to the lack of personal experience (Vrij [2008]). Dissociation might induce greater use of group references rather than self-references. Accordingly, liars are less immediate than truth-

[8]Response length is highly positively correlated with our measure of lexical diversity defined as the number of distinct words. As a result, we include only response length in our analysis.

[9]http://wordnet.princeton.edu/

tellers and refer to themselves less often in their stories (Newman et al. [2003]). Similarly, Bachenko et al. [2008] argues that deceptive statements may omit such references entirely. Regarding references to others, Knapp et al. [1974] find that deceivers typically use more references to other people than truth-tellers; whereas Newman et al. [2003] find the opposite result.

Consistent with prior literature, we hypothesize that deceptive executives have fewer self-references (I) and more first-person plural pronouns (we) in their narratives. Consistent with prior studies, third-person plural pronouns (they) have ambiguous association with deception. We also use the impersonal pronouns (ipron) category, which includes words related to general statements (such as everybody, anybody, and nobody), as an indicator of deception. Although the association with deception is theoretically ambiguous, prior research generally finds that deceivers use more generalizations. We also hypothesize that deceptive statements include more references to general (or audience) knowledge in order to gain credibility. We construct a new word category to measure the use of references that is references to general knowledge (genknlref), which includes phrases such as "you know", "others know well", etc.

Next, negative statements are recognized as indicators of a deceptive message (e.g., Adams and Jarvis [2006]). Accordingly, Vrij [2008] argues that lies often include statements indicating aversion towards a person or opinion, such as denials and statements indicating a negative mood. To capture this dimension, we use the LIWC categories of negation, anxiety, swear words, anger, assent, and positive emotions. We hypothesize that negation, anxiety, swear words, and anger are positively related to deceptive statements, whereas assent and positive emotions are negatively related to deception. In addition, for positive and negative emotions words, we differentiate between "extreme" and "non-extreme" words. In particular, we hypothesize that in our setting executives are using extreme positive emotions words

such as "fantastic" to sound more persuasive while making a deceptive claim.[10] We expect to observe a positive association between extreme negative emotions words and deception.

According to the lack of embracement perspective, liars lack conviction and differ from truth-tellers on the degree of certainty in their statements. To reinforce this concept, previous studies (e.g., Adams and Jarvis [2006], Bond and Lee [2005], Newman et al. [2003]) argue that tentative words imply distance between the speaker and his/her statements. Hence, we expect a positive relation for tentative (tentat) words and a negative relation for words that connote certainty (certain) with deception.

Finally, based on our reading of ten likely deceptive transcripts, we develop two categories "shareholders value" (includes phrases such as "shareholder welfare", "value for investors", etc.) and "value creation" (includes phrases such as "creates value", "unlocks value", etc.) and expand the LIWC list of hesitations. Similar to the discussion above, there is ambiguity about the direction of association with deception for these categories. Specifically, according to the cognitive effort perspective, liars should use more hesitation words; whereas according to the control perspective, liars should use fewer hesitation words due to preparation. Similarly, if "shareholders value" and "value creation" categories capture the general nature of statements made by executives, we would expect a positive relation with deception. However, consistent with the control perspective, liars can consciously avoid the use of these categories in order to mitigate the personal consequences that can arise from shareholder lawsuits after the discovery of accounting malfeasance. Under this scenario, we would expect a negative relation between the use of "value" statements and deception.

[10]To construct both categories of extreme positive and negative emotions words, we selected the words that in our opinion express strong emotions from correspondingly posemo and negemo LIWC categories and completed the lists by adding synonyms for these words from WordNet.

4. Sample

We construct our sample using a comprehensive set of conference call transcripts provided by FactSet Research Systems Inc.[11] We consider all available transcripts of quarterly earnings conference calls for the U.S. companies over the time period from 2003 to 2007. The transcripts are in *.xml* format (an example of the Earnings Conference Call for the first quarter of 2007 for Staples Inc. is presented in Appendix A). A total of 29,663 transcripts were automatically parsed.

As illustrated in the Appendix A.1 (*.xml* file), the typical conference call consists of a Management Discussion section and a Question and Answer (Q&A) section. Our discussions with investor relations professionals indicate that Management Discussion section is thoroughly prepared and screened by legal and investor relations departments, and, thus, we believe that this text is not a desirable data source for detecting executive deception. In contrast, although there is likely to be some preparation provided to CEOs and CFOs for expected questions from analysts, the Q&A section is more likely to be spontaneous. Since spontaneous speech in the Q&A section has more potential to provide insights into executive deceptive behavior, this text (or corpus) is our primary data source.

The transcript of a conference call is generally well structured, and this enables us to automatically extract the necessary data for the linguistic analysis . The first part of a file contains names of corporate representatives, outside participants, and their speaker identifiers. In addition, transcripts have an operator (who coordinates the call) with his/her own identifier. There are three types of phrases that can be found in Q&A sections: operator's introductory phrase, answer, and question. In general, each speaker has an identifier and a type of the phrase that belongs to him/her (question or answer). We assume that all answer phrases belong to corporate representatives and all question phrases belong to

[11] See: http://www.factset.com/data/factset_content/callstreet

outside speakers. In order to identify each speaker, it is necessary to know his/her specific identifier. However, speaker identifiers are not provided consistently, and we make several assumptions in our parsing algorithm. Since the operator introduces each new outside participant, we assume that the same participant keeps asking questions until operator introduces another participant. Further, because the operator does not typically introduce corporate representatives at the Q&A section of a conference call, we assume that the same corporate representative continues to answer questions until a new corporate representative is identified.

Our parsing algorithm of *.xml* files involves the following: (i) an operator phase precedes the first question in a session, otherwise, questions and answers are not recorded, (ii) the same speaker keeps asking questions until operator interrupts, (iii) the same speaker keeps answering questions until new speaker who also answers questions interrupts, and (iv) a question must come first after operator speaks. This procedure produces a database where we can track the question posed by a speaker and the answer from a corporate representative that follows after a particular question.

We define an instance as all answers of a corporate representative (e.g. CEO, CFO, COO, etc.) at a particular conference call. For example, Appendix A.2 presents the records in the database that correspond to the Q1 2007 conference call of Staples Inc. There were five people answering questions at Staples's conference call. Moreover, from the header of *.xml* file, we can find title of a corporate representative. As an example of an instance, we present the partial text for JOHN MAHONEY (the CFO of Staples Inc.) in Appendix A.3.[12]

Since CEOs and CFOs are the most likely to know about financial statement manipulation, and these executives are the most common participants on the conference call, we

[12]One undesirable feature of the conference call is that the names for corporate individuals can be written differently on the same transcript, and each different name is given its own speaker id. For instance, BEDI AJAY SINGH can be called BEDI SINGH or EDWARD PARRY as ED PARRY or RICHARD NOTEBAERT as DICK NOTEBAERT and so forth. To achieve better accuracy in compiling all instances of the same person at a particular conference call into one instance, we manually correct these inconsistencies.

develop separate data file for CEOs and CFOs. We constrain the length of an instance to be greater or equal to 150 words which corresponds approximately to an answer to one question. Our CEO sample has 16,577 instances and CFO sample has 14,462 instances.

We also develop another sample (DAC sample) which combines instances for CEO/CFO and the necessary accounting variables for the estimation of the model with discretionary accruals. To construct the DAC sample, we take the overlap of the CEO and the CFO sample and require non-missing values for the computation of discretionary accruals and other control variables (described below). The DAC sample has 5,181 instances.

The descriptive statistics for our samples are presented in Table 2.[13] Approximately 90% of our firms are listed on NYSE or NASDAQ (Panel A). We find that industry distribution in our sample is close to the Compustat industry distribution (Panel B). Our sample is also significantly larger in terms of market capitalization, total assets, and sales. Further, firms from our sample are more profitable in terms of ROA and profit margin, have significantly greater free cash flows, and equivalent sales growth relative to the Compustat population.

5. Methodology

5.1 Measurement issues

In studying verbal cues of deception, previous research often uses controlled experiments where participants are asked to lie or to tell the truth (e.g., Newman et al. [2003], Bond and Lee [2005], Hobson et al. [2010]). This design allows certainty about whether a statement is deceptive or not, but the somewhat contrived nature of this type of experiment can differ immensely from real world lying and result in serious threats to external validity. In contrast, we analyze a real world setting where we know that the quarterly financial statements discussed by the CEO and CFO during the conference call were subsequently

[13] We only present descriptive statistics for the middle year of our sample, 2005, in order to be parsimonious. The descriptive statistics are comparable for the other years.

restated. We assume that these executives either intentionally manipulated the financial reports or that they have knowledge that they are providing investors with false information during the call.[14]

We use data from Glass, Lewis & Co. to identify quarterly reports that are restated by each firm. These data cover restatements announced during the time period from 2003 to 2009. In order to identify serious restatements, as opposed to "trivial" restatements, we require deceptive conference calls to exhibit a material weakness, a late filing, an auditor change, or a disclosure using Form 8-K. A material weakness implies that there is a deficiency in the internal controls over financial reporting that can make it easier for executives to manipulate. An auditor change can be a signal about deficiency in monitoring. A late filing implies that it takes time for a firm to correct the accounting, which suggests that the manipulation is complex and possibly intentional. Finally, Plumlee and Yohn [2008] show that a Form 8-K filing is related to more serious restatements.

We conjecture that verbal cues of deception are more likely to be observed when the restatements are relatively large. Our measure for the size of manipulation is computed by accumulating the bias in reported net income over all restated quarters in Glass, Lewis & Co. data.[15] We compute the bias for a given quarter as the difference between net income originally reported for the fiscal quarter and the latest value of net income available in Compustat Point In Time History table. Point In Time History table has an additional time dimension for each quarter end date (datadate) that identifies the end of the month

[14]We acknowledge that our approach to labeling a transcript as deceptive has measurement error related to whether the conference call participants actually know about the deception. To mitigate this concern, we impose several criteria for labeling transcripts as deceptive, which reflect different degrees of restatement seriousness. In addition, our statistical analysis is likely produce conservative results because some manipulated quarters are never restated or restated outside of the time period we examine.

[15]For instance, for a hypothetical firm that files two restatements for three quarters (e.g., one restatement is for Q1 2002, and the second restatement is for Q1 2001 and Q3 2002), we define the corresponding bias for each restated quarter as $Bias_{Q1\ 2001} = dNIQ_{Q1\ 2001}$, $Bias_{Q1\ 2002} = dNIQ_{Q1\ 2001} + dNIQ_{Q1\ 2002}$, $Bias_{Q3\ 2002} = dNIQ_{Q1\ 2001} + dNIQ_{Q1\ 2002} + dNIQ_{Q3\ 2002}$, where $dNIQ_t$ is the difference between net income originally reported for fiscal quarter t and the latest value of net income available in the Compustat Point In Time data for this quarter.

that the data value is the part of Compustat (pointdate) that has maximum spread of 60 months.[16] Specifically, for every quarter, we take a value of net income that corresponds to the earliest pointdate (i.e., the value that is originally reported) and as the latest value of net income we take the value that corresponds to the latest pointdate (i.e., the value for that quarter net income that is last available in Compustat according to Point In Time History table). If net income is restated then the original value differs from the latest available value. Following Palmrose et al. [2004], we scale cumulative measure of the bias in net income by the originally reported total assets.

In order to provide insight into whether linguistic features of deception vary with the size of a restatement, we separate instances into several categories according to a set of criteria summarized in Appendix B.[17] These criteria are the no-threshold (NT) criterion that ignores the magnitude of the bias, the absolute value of bias criteria for the bias that is greater than 25th (AS25) and 50th (AS50) percentiles of the non-zero absolute value of bias distribution, and the positive value of bias criteria for the bias that is greater than 25th (PS25) and 50th (PS50) percentiles of the non-zero positive value of bias distribution. We interpret deception as any deviation from the truth (i.e., the sign of the bias should not matter), and thus we primarily focus on the NT, AS25, and AS50 criteria. We discuss the results obtained from other deception criteria in Section 7.

The frequency of deceptive firm-quarters (labeled as deceptive under the NT, AS25, and AS75 criteria) by year is presented in Table 3. Years 2003, 2004 and 2005 have the highest rate of deceptive firm-quarters. This result is likely due to the fact that there is more time after the accounting manipulation for detection.[18] As should be expected, the overall

[16]See: Overview of COMPUSTAT Preliminary, Unrestated and Point-in-Time Datasets on WRDS.

[17]We could label as deceptive only restatements that are related to SEC enforcement actions or security litigation. However, this considerably restricts the number of deceptive transcripts. Moreover, SEC enforcement actions and security litigation appear to be subject to a selection bias (e.g., Correia [2009])

[18]This observation highlights that there is likely to be more measurement error in assigning instances to the deceptive category for 2006 and 2007. This measurement problem will reduce the power of our statistical analysis, and thus produce conservative statistical tests.

percentage of deceptive firm-quarters is the highest for the less restrictive NT criterion which is 13.59% (14.20%) for the samples of CEOs (CFOs) and the lowest for the most restrictive AS50 criterion which is 4.98% for the both samples of CEOs and CFOs (Table 3).

In order to build a classification model, we convert each instance into a vector in the space of word categories summarized in Table 1. To construct word-based variables, we compute a number of words present in the transcript of Q&A in each category. Similar to most of the prior literature, we assume that an instance is just a "bag-of-words" (i.e., the position of a word in a sentence is irrelevant for classification, and "context" is ignored). To count words in word categories specified in Table 1, we use R text-mining package `tm` (Feinerer et al. [2008], Feinerer [2010]). We divide the counts by the total number of words in the instance (instance length) and multiply by the median instance length in the sample. This procedure standardizes word counts in such a way that a unit increase in the standardized word count corresponds to a one word increase in the document of the sample-specific median length.

Descriptive statistics for the word-based variables for the samples of CEOs and CFOs are presented in Table 4. CEOs have much longer instances than CFOs with the mean (median) instance length for CEOs of about 1,811 (1,611) words and the mean (median) instance length for CFOs of about 987 (777) words. Both CEOs and CFOs have impersonal pronouns as the largest category in References with references to general knowledge having the lowest word count. The largest category for Positives/Negatives is non-extreme positive emotions words with negations being the second largest category. As might be expected, the category of swear words has the lowest count. Both executives use almost twice as many tentative words as words expressing certainty. There ere are very few hesitations, shareholders value, and value creation words in transcripts' Q&As.

5.2 Econometric issues

Similar to traditional classification research, we estimate a simple binomial logistic model. The outcome variable is coded as one if a conference call is labeled as deceptive and zero otherwise. To estimate the prediction error of a classifier, it is necessary to estimate the out-of-sample prediction error, because the in-sample prediction error is very optimistic estimate of the prediction error on a new data set. One approach is to randomly split the sample into two parts, and use one part to estimate the model and the other part to obtain the out-of-sample prediction error using the estimated model. However, deceptive outcomes are rare events and single split may not provide enough variation to fit the model and to consistently estimate the out-of-sample prediction error.

To obtain a consistent estimate of the prediction error, we perform cross-validation which is generally recommended for finite samples (Efron and Tibshirani [1994], Witten and Frank [2005], Hastie et al. [2003]). Specifically, the K-fold cross-validation is implemented in the following manner: (1) data is split into K roughly equal samples (folds); (2) $k : k = 1, ..., K$ fold is fixed; (3) the model is estimated using $K - 1$ folds ignoring the kth fold; (4) performance of the model is evaluated using the kth fold. These steps are repeated K times where the $k = 1, .., K$. Although, there is no theoretical justification for a particular number of folds K, 10-fold cross-validation repeated 10 times is commonly applied in practice (Witten and Frank [2005]). We use this heuristic to estimate our prediction error.

We also implement a stratified cross-validation that implies that the proportion of deceptive and non-deceptive instances in each random data split is the same as in the original sample. Using the cross-validation, we can estimate the mean out-of-sample performance of a particular model and compare across different models. When we compare different models, we evaluate these models using the same split of data. Specifically, we split data first, fix the fold, and then estimate or evaluate different models using the same folds.

Since deceptive instances are rare events, if all instances have equal weight, it can be

optimal to simply classify all deceptive instances as truthful. To deal with this uninformative choice, classification studies typically impose a greater weight on a rare class (Witten and Frank [2005]). In the estimation sample, we impose a unit weight on truthful instances and the greater weight on deceptive instances, which equals to the number of truthful instances divided by the number of deceptive instances.[19] This procedure balances the overall weight placed on truthful and deceptive instances.

There are a number of performance measures that are used in classification studies. The primary performance measures are the accuracy, the true positive rate (TPR), the false positive rate (FPR), and the precision. These measures can be defined using the contingency table that represents the outcomes from a set of predictions. The traditional contingency table is as follows:

		Predicted Class	
		Deceptive	Truthful
Actual Class	Deceptive	True Positives	False Negatives
	Truthful	False Positives	True Negatives

We take deceptive instances as a positive class and truthful instances as a negative class. The True Positives (TP) are the number of deceptive instances classified as deceptive; the False Negatives (FN) are the number of deceptive instances classified as truthful; the False Positives (FP) are the number of truthful instances classified as deceptive; and the True Negatives (TN) are the number of truthful instances classified as truthful. The resulting performance measures of interest are:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \; Precision = \frac{TP}{TP + FP}$$

$$TPR = \frac{TP}{TP + FN}, FPR = \frac{FP}{FP + TN}$$

[19] This weight is about the same as in the original sample because stratified split preserves the proportion of deceptive and non-deceptive instances in the sample.

As discussed in Fawcett [2006], the TPR and FPR can be combined in a Receiver Operating Characteristics (ROC) graph, which is the standard technique for visualizing and selecting classifiers. ROC graphs for two-class problems are two-dimensional graphs in which the TPR is plotted on the y-axis, and the FPR is plotted on the x-axis. If an output of a classifier is the probability of a positive class (as in our binary setting), each probability cutoff for assignment to the positive class will produce a point in the ROC (the TPR and FPR) space. Specifically, each cutoff defines which observations are classified as truthful or deceptive and enables us to build a contingency table to compute the TPR and FPR. By varying cutoffs for the probability of the positive class, it is possible to draw the entire ROC graph. Three important points in the ROC space are $(0, 0)$, which corresponds to issuing a positive classification (i.e., no false positive errors due to no positive assignments), $(1, 1)$ which corresponds to unconditionally issuing positive classifications (i.e., all positives are labeled correctly, and all negatives are erroneously classified), and $(0, 1)$ which corresponds to perfect classification (i.e., all positives are issued correctly and no negatives incorrectly labeled as positives) (Fawcett [2006]).

As explained by Fawcett [2006], random guessing (random classifier) corresponds to the diagonal in the ROC space. For example, if the classifier randomly guesses positives 50% of the time, half of the positives and half of the negatives should be correct, representing the $(0.5, 0.5)$ point in the ROC space. If the random classifier guesses positives 80% of the time, it is expected to get 80% of the positives correctly and 80% of negatives incorrectly producing $(0.8, 0.8)$ in the ROC space. In other words, a random classifier is some point on the diagonal in the ROC space. Therefore, in order to move from the diagonal into the upper triangular region, a classifier must have the ability to correctly predict outcomes by exploiting some information in the data.

ROC graphs do not depend on the class distribution, because the TPR and FPR are row-based ratios. Independence on the class distribution implies that ROC graphs are not affected

by the rarity of positive instances in our study. It is possible to reduce the performance of a classifier to a single scalar by considering the area under the ROC graph (AUC). As discussed in Fawcett [2006], the AUC is equivalent to the probability that a randomly chosen positive instance will be ranked higher by a classifier than a randomly chosen negative instance.[20]

We test the AUC for our classifiers against the AUC for a random classifier with the t-statistic developed using cross-validation. More specifically, we perform 10-fold cross-validation repeated 10 times, which means that we have 100 out-of-sample performances. These 100 observations are used to compute the necessary t-statistics.

To compare the performance of the models that use words or discretionary accruals, we compute the AUC for every out-of-sample run and perform paired t-tests for the significant difference in the means of AUC. We can use paired t-tests, because the classifiers are estimated and tested on the same iteration-specific data splits. This approach enables us to test for differences in AUCs across various models (e.g., the CEO word-based model versus the CFO word-based model or the CEO word-based model versus the CEO word-based model and the discretionary-accruals-based model).

Two other traditional performance measures are the accuracy and the precision. The accuracy is the rate of observations classified correctly, and the precision is the rate of true positives among those observations that are classified as positive. One might argue that the precision is the single most useful performance measure (i.e., we want to identify deceptive observations). However, the precision does not take into account how good the classifier is in detecting positives. The precision only measures how many of the observations classified as positives are really in the positive group. It is possible to have a very high precision when very few of the observations are classified as positive. Thus, the AUC is a more prominent measure for assessing a classifier performance than the precision.

Finally, to test the hypotheses related to specific verbal cues or word categories, we

[20]This probability equals 0.5 for a random classifier that is the area under the diagonal in the unit square.

perform a stratified bootstrap to obtain confidence intervals (Efron and Tibshirani [1994]). In particular, we draw (with replacement) a random sample of the same size and proportion of truthful and deceptive instances as the original sample and estimate coefficients of the logistic regression. We perform 1,000 bootstrap replications and report 10%, 5% , 1% significance (two-tailed) confidence intervals.[21]

6. Results

6.1 Financial firms

Since our methodology is somewhat new to accounting research, we first present results for a sample of firms from the financial sector. This example will enable us to discuss the results in a simple manner. The sample of financial firms is also interesting, because typical discretionary accrual models are problematic to apply for this important group of firms. We classify firms as financial following Global Industry Classification Standard (Compustat GSECTOR is 40). We present results only for the no-threshold (NT) criterion as the number of deceptive instances under the absolute value of bias criteria (AS25, AS50) is below 50, and it is difficult to produce reliable conclusions with such a small number of deceptive instances.

The ROC curve for the sample of financial firm CEOs is presented in Figure 1. Each point on the ROC curve is the average TPR and FPR for a fixed threshold over 100 classifications from 5-fold cross-validation repeated 20 times.[22] The bars on the curve show the 95% confidence intervals at the thresholds equal 0.25, 0.50, and 0.75. As the ROC exhibits more curvature, more deceptive instances can be classified correctly (TPR) at the cost of classifying fewer non-deceptive instances incorrectly as deceptive (FPR).

[21]For this paper we extensively employed R software (R Development Core Team [2005]) and specific packages: glmnet (Friedman et al. [2009]), boot (Davison and Hinkley [1997], Canty and Ripley [2009]), ROCR (Sing et al. [2005]), xtables (Dahl [2009]), and tm (Feinerer et al. [2008], Feinerer [2010]).

[22]We use 5-fold cross-validation repeated 20 times for financial firms to have larger number of deceptive instances in our testing samples.

Classification models based on the word categories for CEOs (CFOs) perform significantly better than a random classifier that does not use any information (Table 5). In particular, the AUC for CEOs is 58.93%, which is significantly better than 50% (the AUC for a random classifier). The AUC for CFOs is less impressive, 53.56%, however, significantly better that 50%. Overall accuracy (the percentage of calls classified correctly) for both CEOs and CFOs is higher than 60%. At the same time, the model that classifies all conference calls as non-deceptive would achieve the accuracy of around 90%, but would not perform significantly better than any other random classifier.

We also observe that several word categories have a significant association with the likelihood of deception. The references to general knowledge, impersonal pronouns, words expressing extreme negative emotions, and words of assent are associated with deception in predicted direction (Table 5). However, tentative and certainty words have unexpected signs. To simplify interpretation, Table 5 reports factors by which the odds of a deceptive conference call should be multiplied if the number of words in the category increases by 1% of the median instance length. Specifically, the word categories that increase (decrease) the likelihood of deception have factors that are greater than (less than) one.[23] More specifically, if there are two CEO instances that differ only in the number of words from references to general knowledge category, the instance with 15 (1% of 1,526, i.e. the median length of instances of financial firm CEOs) more words in this category is 1.69 times as likely to be deceptive as the other one.

6.2 The no-threshold and the absolute value of bias criteria

For the NT, AS25, and AS50 criteria, classification models based on verbal cues of CEOs and CFOs perform significantly better than a random classifier by about 4%-6% (Table 6). The

[23] Odds refer to the ratio of the probability of a deceptive call to the probability of a non-deceptive call. To illustrate, let probability of a deceptive conference call be 10% then the odds of a deceptive call is $.10/(1-.10) = 1/9$ or 1 to 9.

overall accuracy is approximately 50% - 65%, and the percentage of calls that are deceptive and classified as deceptive ranges from about 6% under the most restrictive AS50 criterion to about 17% for the least restrictive NT criterion.

These classification performance measures are conservative for two reasons. First, the AUC, the precision, and the accuracy are obtained using a 10-fold cross-validation repeated 10 times, when we compute these measures out-of-sample. The out-of-sample performance is conservative relative to the performance computed from estimating and testing the model on the same set of data. Second, there can be some manipulations that are never revealed or are revealed later in time. Consequentially, some of the non-deceptive conference calls are actually mislabeled. It can be the case that our model correctly classifies these observations as deceptive in the testing set, but this classification would be marked as incorrect lowering the accuracy, the precision, and the AUC measures.

The estimated associations between verbal cues and the likelihood of deception for CEOs (Panel A, Table 7) and CFOs (Panel B, Table 7) are mostly consistent with prior theory and empirical deception research.[24] Surprisingly, the signs of association for some word categories differ for CEOs and CFOs. These are third person plural pronouns and certainty words. Whereas the prior research finds both positive and negative relationship between third person plural pronouns and the likelihood of deception, the sign for certainty words is expected to be negative. However, deceiving CEOs use fewer certainty words, which is consistent with the theory, but deceiving CFOs use more certainty words, which contradicts the theoretical predictions.

Some categories are significant only for one executive (Table 7). Deceiving CEOs use fewer self-references, more impersonal pronouns, more extreme positive emotions words,

[24]We report the estimates for the factors by which the odds of a deceptive instance should be multiplied if the number of words in a category increases by 1% of median instance length except for word count (for which the factor corresponds to the increase in the length of an instance by the median instance length) and for rare categories such as swear words, hesitations, and references to shareholders value and value creation (for which the factor corresponds to the increase in the number of words in the corresponding category by one word). The factor is greater than (less than) one for a coefficient which is greater than (less than) zero.

fewer extreme negative emotions words, and fewer hesitations. The use of fewer hesitations by CEOs in deceptive calls can be the consequence of CEOs having more prepared answers or answering planted questions. Similar to the results for extreme negative emotions words, CEOs use fewer swear words, which is inconsistent with our theoretical prediction. However, for CFOs the use of extreme positive emotions and extreme negative emotions words is not significantly associated with deception. In contrast to CEOs, CFOs use significantly more tentative words.

Various word categories are significantly related to the probability of deception for both CEOs and CFOs. Both executives have larger number of words in deceptive answers, although the increase in the number of words would need to be very large in order to alter the odds of a deceptive call in a substantial way. For example, increasing the length of a CEO's instance by the median instance length (i.e., 1,611 words) increases the odds of a deceptive instance by a factor of 1.10 under AS50 criterion. Another category that is consistent throughout different criteria is the references to general knowledge. Both CEOs and CFOs have more words that reference general knowledge such as "you know" in deceptive instances. Deceptive calls also have fewer non-extreme positive emotions words and mention shareholders value and value creation less often.

Although majority of word categories are associated with the likelihood of deception in the predicted direction, there are some notable exceptions. Specifically, first person plural pronouns exhibit both positive and negative associations with deception, words for assent have an unexpected significant positive association, and words that express anxiety an unexpected significant negative association with deception. Finally, words that express negation and anger are not significantly related to deception.

The interpretations above are based on the word categories that have a statistically significant association with the likelihood of deception at 10% significance level (two-tailed) for at least one of the criteria. However, some categories are significantly related to deception

under all three criteria. For the sample of CEOs, these are anxiety words, extreme negative emotions words, and references to value creation. For example for CEOs, the increase by 1% of median instance length in extreme negative emotions words decreases the odds for a call to be deceptive by a factor of 0.76 for NT, 0.67 for AS25, and 0.63 for AS50. In contrast, the increase by one word in the words related to value creation decreases the odds of a call to be deceptive by a factor of 0.75 for NT, 0.65 for AS25, and 0.67 for AS50.

The word categories significant under all three criteria differ for the sample of CFOs. These are references to the general knowledge and non-extreme positive emotions. An increase by 1% in the number of references to general knowledge increases the odds of a call to be deceptive by a factor of 1.39 for NT, 1.18 for AS25, and 1.14 for AS50. Finally, the increase by 1% in the number of words expressing non-extreme positive emotions decreases the odds of a call to be deceptive by a factor of 0.92 for NT, 0.90 for AS25, and 0.82 for AS50.

To estimate the likelihood of deception in future studies, we propose two models (for CEOs and for CFOs). These models are for detecting narratives which are deceptive under NT criterion, which has the largest number of calls labeled as deceptive, and include coefficients significant at 1% and 5% significance level only. Specifically, the logistic model for predicting the probability of a deceptive narrative ($Y_i = 1$) using word categories X_i:

$$Pr(Y_i = 1|X_i) = \frac{1}{1 + e^{-f(X_i)}},$$

where linear predictor $f(X_i)$ differs for CEOs and CFOs[25]:

$$\begin{aligned} f(X_i)^{CEO} &= 0.678 - 0.004 \cdot I - 0.004 \cdot we + 0.025 \cdot genknlref - 0.005 \cdot posemone \\ &+0.013 \cdot posemoextr - 0.156 \cdot swear - 0.017 \cdot negemoextr - 0.290 \cdot value \\ f(X_i)^{CFO} &= 0.0001 \cdot wc - 0.019 \cdot they + 0.047 \cdot genknlref + 0.015 \cdot assent \\ &-0.013 \cdot posemone - 0.058 \cdot anx + 0.011 \cdot certain - 0.909 \cdot shvalue \end{aligned}$$

6.3 Models with verbal cues vs. model with discretionary accruals

Although our models perform better than a random classification, it is important to also compare this performance to more traditional models used in accounting research. We implement this comparison by using a classification model based on discretionary accruals plus other relevant control variables Correia [2009].[26]

To compute discretionary accruals, we follow methodology of Kasznik [1999]. Specifically, for each two-digit SIC code and every fiscal quarter we estimate the model for total accruals (Table 8):

$$TAC_{it} = \beta_0 + \beta_1(\Delta Sales_{it} - \Delta AR_{it}) + \beta_2 PPE_{it} + \beta_3 \Delta CFO_{it} + \epsilon_{it},$$

where TAC_{it} is the measure of total accruals computed as in Hribar and Collins [2002] for firm i and fiscal quarter t; $\Delta Sales_{it} - \Delta AR_{it}$ is the change in revenues adjusted for the change in receivables; PPE_{it} is gross property, plant, and equipment; ΔCFO_{it} is the change in operating cash flows. All variables are deflated by the mean of total assets (Table 8,

[25]Note that the suggested models use the actual estimated coefficients significant at 1% and 5% (two-tailed); whereas in tables, we report factors by which the odds of a deceptive instance are changed. Specifically, in tables we report factors $f = e^{N\beta}$, where N is the number of words and β is the estimated coefficient. Hence, $\beta = ln(f)/N$.

[26]Most prior studies that use discretionary accruals to measure the extent of earnings manipulation use annual data. However, we use quarterly data to be consistent with the frequency of quarterly earnings conference calls. As a consequence, our results can differ from prior literature because the discretionary accrual model might not be applicable for quarterly data.

Panel A). We estimate the regression for the total accruals by the two-digit SIC and the fiscal quarter and require at least 40 observations for estimation. We use Compustat Point In Time Historical table, which, in contrast to Compustat Quarterly, has originally reported financial items. Following the prior literature, we exclude financial services industry (SIC codes 6000-6999) and utilities (SIC codes 4900-4999). We compute discretionary accruals as a residual from the regression for the total accruals.

Our model for predicting deceptive conference calls is based on the logistic model of Correia [2009], which uses the absolute value of discretionary accruals, the actual security or debt issuance, the market capitalization, the free cash flows, and the growth in cash sales to predict restatements from Glass, Lewis & Co. data. To be consistent with the regression for the total accruals, we deflate the market capitalization and the free cash flows by the average total assets. We refer to the model that includes only accounting variables as the discretionary accruals (DAC) model. The only measure which is consistently significantly associated with the deception is the free cash flow variable, where higher values are associated with lower probability of deception.

We compare the performance of six different models (CEO, CFO, DAC, DAC+CEO, DAC+CFO, DAC+CEO+CFO). These pairwise comparisons are based on 10-fold cross-validation repeated 10 times. In order to mitigate the noise introduced by using different estimation and testing samples across models, we estimate and test the six models using the same data for estimation and the same data for testing. That allows us to use paired t-tests to test for the difference in AUC measures (Table 9).

Under the NT criterion, we find that models that include only word categories for CEOs or CFOs have the significantly higher AUC than the DAC model. For the AS25 criterion, the only word categories for CEOs model performs statistically worse than the DAC model. However, the model that includes word categories only for CFOs has the AUC that is statistically equivalent to the DAC model. We find similar results for the AS50 criterion, the CFO

model has the AUC significantly higher than that of the DAC model, whereas the CEO model is statistically equivalent to the DAC model. As should be expected, the DAC+CEO+CFO model has the significantly higher AUC than that of the DAC model under all three criteria. Thus, we find that the linguistic variables (especially for the CFO) add the incremental classification power to the DAC model (or that the accounting model adds the incremental classification power to the linguistic model).

7. Extensions

7.1 The positive value of bias criteria

The results in Table 7 are developed ignoring whether accounting manipulations increased or decreased reported earnings. Some researchers and analysts argue that the positive bias in reported earnings can be more harmful for investors, as it can cause prices to be higher than justified by the real earnings figure. It is also quite possible that the word categories that predict deception involving the positive bias can differ substantially from the word categories that predict the absolute value of bias. For instance, executives may use more positive emotions words and fewer negative emotions words when they bias net income upwards. In Table 10, we provide the performance results from our linguistic models that use the PS25 and PS50 criteria. Models based on the word categories predict deception involving the positive bias better than a classifier that uses no information by 3%-5%. Similar to the results in Table 6, overall accuracy is about 50% - 60%. As expected, the percentage of the deceptive instances among those classified by the model as deceptive (precision) under the PS25 (PS50) criteria is lower than the model precision under the AS25 (AS50) criteria. The precision is reduced mechanically because the proportion of deceptive calls under the positive bias criteria is lower. As a result, the precision of classifying deceptive calls is approximately 4% (PS50) to 6% (PS25) in Table 10 compared to 6%(AS50) to 8% (AS25) in Table 6.

With respect to the logistic models (not reported), most coefficients are similar to the coefficients for the absolute value of bias criteria. However, there are a number of differences. First, the word category of negations is insignificant for the absolute value of bias criteria, but has a significant negative association with deception for the positive bias criteria for both CEOs and CFOs. This result is not consistent with prior theoretical explanations. Second, extreme negative emotions words become insignificant, whereas this category has a significant negative relation to deception for CEOs. Finally for CEOs, some categories become significant for more serious criteria such as the self-references (PS50) and references to shareholders value categories (PS25 and PS50) have a negative association with deception, whereas the extreme positive emotions category (PS25) has a positive association with deception.

7.2 Individual fixed effects

There is the possibility of spurious classification results based on the composition of our sample and the way we perform cross-validation. Approximately 70% of firms that have at least one deceptive call have more than one deceptive call (i.e., there is some clustering of deceptive instances among particular individuals). At every run of the 10-fold cross-validation procedure, we split the sample randomly. The presence of several instances of deception for the same executive implies that for him/her some instances of deception can be in the estimation sample and other instances of deception can be in the test sample. A problem arises when the style of communication and word choice is individual-specific and persistent over time (e.g., some individuals are overall positive and some are overall negative). As a result, the correct classification in the testing sample may come from the fact that we have the same individual deceiving in the estimation sample. That is the individual-specific fixed effect is the reason for correct classification and not a pattern of deception.

Another related issue occurs when we estimate the prediction model using the unadjusted

counts for word categories. Specifically, our prior results assume that there is a common benchmark for truthful and deceptive language for all individuals in our sample. However, it is perhaps more reasonable to assume that each individual has his/her own mode for truthful and for deceptive language. This suggests that we can improve the classification performance by adjusting simple word counts for individual fixed effects. Although this seems like a reasonable approach, it is important to realize that it will be difficult to develop a good estimate for a normal (truthful) word count benchmark for an individual because our time series is somewhat short. When we adjust word counts by the average word counts over all previous quarters (requiring the minimum of two quarters) including both deceptive and non-deceptive instances, we find that the classification power becomes very weak.

However, when we adjust word counts by the average over all previous quarters (requiring the minimum of two quarters) excluding deceptive instances, we find that there is a significant classification power in the model (Table 11). Under this adjustment, the number of deceptive quarters is not monotonic across the NT, AS25, and AS50 criteria. For example, the NT criterion has the lowest number of deceptive instances for both CEOs and CFOs in the adjusted setting because it has the biggest number of deceptive instances in the unadjusted setting. Thus, the NT criterion has the smallest number of instances available for adjustment resulting in smaller number of adjusted ones. The logistic models that use the adjusted word categories have the AUC that is 5% - 10% significantly greater than the AUC for a random classifier. Although under NT and AS50 there is a decrease in the AUC using the adjusted word categories for CEOs, we find that there is substantial performance improvement using the adjusted word categories for CFOs. These suggest that our classification results using the unadjusted word categories are unlikely to be entirely spurious.

Most of the coefficients using the adjusted word categories that have the same sign of association with deception are of similar magnitude as the coefficients of the unadjusted categories (Table 12). However, there are a number of substantive changes for both the

CEO and CFO samples. The most significant change is the change in the sign of association between the references to general knowledge and deception. There is a positive association with the likelihood of deception when this category is unadjusted (Table 7) and a negative association with the likelihood of deception when we adjust it (Table 12).

In the sample of CEOs, the assent category changes its sign to the one predicted by theory, whereas the extreme negative emotions and anxiety categories become insignificant. In contrast, words expressing certainty become positively associated with the likelihood of deception, which is inconsistent with theory, but tentative words become significantly positively associated with deception in line with theory. Under the NT criterion, hesitations and shareholder value related words have a positive relation with deception. However, under the more serious AS50 criteria, there is a negative association between hesitations and deception. One speculative interpretation of these results is that the more serious (and practiced) effort is invested into rehearsing presentations involving more substantial lies.

Similarly for the sample of CFOs, the assent category changes its sign to the one predicted. In addition, self-references become negatively associated with deception, and impersonal pronouns become positively associated, which is consistent with theory. Similarly, swear words now have the expected positive sign. In addition, extreme positive and negative emotions words become significantly related to deception with the adjusted categories, when they are not significantly related to deception with the unadjusted categories. In particular, the extreme positive emotions category has negative sign and the extreme negative emotions category has positive sign. There is a divergence in signs for the non-extreme positive emotions and value creation categories for the sample of CFOs between the NT and AS25 (AS50) criteria. In particular, under AS25 and AS50 criteria the sign of association is as expected: negative for the non-extreme positive emotions and value creation categories. One explanation for these results is that the linguistic model may work better when the lies are more substantial.

8. Concluding Remarks

Considerable accounting and finance research has attempted to identify whether reported financial statements have been manipulated by executives. Most of these classification models are developed using accounting and financial market explanatory variables. Despite extensive prior work, the ability of these models to identify accounting manipulations is modest.

In this paper, we take a different approach to detecting financial statement manipulations by analyzing linguistic features present in CEO and CFO remarks during quarterly earnings conference calls. Based on prior theoretical and empirical research from psychology and linguistics on deception detection, we select the word categories that theoretically should be able to detect deceptive behavior by executives. We use these linguistic features to develop classification models for a very large sample of quarterly conference call transcripts.

A novel feature of our methodology is that we know whether the financial statements related to each conference call were substantially restated in subsequent time periods. Since the CEO and CFO are likely to know that financial statements have been manipulated, we are able to identify which executive discussions are actually "deceptive". Thus, we can estimate a linguistic classification model for detecting deception and test the out-of-sample performance of the model.

We find that our linguistic classification models based on CEO or CFO narratives perform significantly better than a random classifier by 4% - 6% with the overall accuracy of 50% - 65%. In terms of linguistic features of the narratives, we find that both CEOs and CFOs use more references to general knowledge, fewer non-extreme positive emotions words, fewer shareholders value and value creation references. We also find that the pattern of deception for CEOs differs from the pattern on deception for CFOs. Specifically, CEOs use fewer self-references, more third person plural and more impersonal pronouns, fewer extreme negative emotions words, more extreme positive emotions words, fewer certainty words

and fewer hesitations. In contrast, CFOs do not use more extreme negative and extreme positive emotions words. Finally, we find that linguistic features statistically improve the out-of-sample performance for a traditional accounting-based model that uses discretionary accruals. These performance results suggest that it is worthwhile for researchers to consider linguistic cues when attempting to measure the quality of reported financial statements.

As with any exploratory study, our findings are subject to a number of limitations. First, we are not completely certain that the CEO and/or CFO know about the manipulation when they are answering questions during the conference call. This issue will cause our deception outcome to be measured with error. Second, simply counting words ("bag-of-words") ignores important context and background knowledge. Third, we use general psychosocial dictionary, LIWC, which may not be completely appropriate for capturing business communication. Finally, although we have a large comprehensive set of conference calls, our sample consists of relatively large and profitable firms. This limits our ability to generalize our results to the population of firms.

In terms of future research, it would be useful to refine general categories to business communication.[27] It would also be desirable to adapt natural language processing approaches to capture the context of word usage for identifying deceptive executive behaviors. Finally, it would be interesting to determine whether portfolios formed on the basis of our word-based measure of deception generate future excess returns (alpha) and/or help eliminate extreme losers from a portfolio selection.

[27] An alternative lexical database is WordNet (http://wordnet.princeton.edu/).

Appendix A. Conference call of Staples Inc., Q1 2007

A.1. *.xml* file

```
<?xml version="1.0" encoding="us-ascii" ?>
- <transcript id="1079826" product="CorrectedTranscript" xmlns="http://www.factset.com/callstreet/xmllayout/v0.1">
    - <meta>
      <title>Q1 2007 Earnings Call</title>
      <date>2007-05-22</date>
    - <companies>
        <company>472</company>
      </companies>
    - <participants>
    <participant id="0" type="operator">Operator</participant>
    <participant id="1" type="corprep">Laurel Lefebvre, Investor Relations</participant>
    <participant id="2" type="corprep">Ronald L. Sargent, Chairman and Chief Executive Officer</participant>
    <participant id="3" type="corprep">Michael Miles, President and Chief Operating Officer</participant>
    <participant id="4" type="corprep">John J. Mahoney, Vice Chairman and Chief Financial [...]</participant>
    <participant id="5">Matthew Fassler</participant>
    <participant id="6">Ronald Sargent</participant>
        ...
        </participants>
      </meta>
- <body>
  - <section name="MANAGEMENT DISCUSSION SECTION">
    - <speaker id="0">
      - <plist>
        - <p>
            Good day ladies and gentlemen and welcome to the First Quarter 2007 Staples
```

```
            Incorporated Earnings Conference Call [...]
          </p>
        </plist>
      </speaker>
    - <speaker id="1">
      - <plist>
        <p>Good morning everyone and thanks for joining us for our first quarter 2007
         earnings announcement. During today's call [...]
         </p>
        ...
       </plist>
      </speaker>
    </section>
- <section name="Q&A">
    - <speaker id="0">
      - <plist>
        - <p>
          <mark type="Operator Instructions" />
          . And your first question comes from the line of Matthew Fassler with Goldman Sachs.
          </p>
          </plist>
      </speaker>
      - <speaker id="5" type="q">
        - <plist>
            <p>Thanks a lot and good morning.</p>
          </plist>
        </speaker>
      - <speaker id="6" type="a">
```

```
  - <plist>
      <p>Good morning Matt.</p>
    </plist>
  </speaker>
- <speaker id="5" type="q">
  -<plist>
     <p>I've got two quick related questions for you. As we look at your guidance,
     you didn't really change the stated sales guidance. So, to the extent that you
     are directing the numbers towards the low end of the range, [...] or is that a
     manifestation in your view of perhaps having to push a little harder somewhere
     on the margin front in order to capture the sales?
     </p>
   </plist>
  </speaker>
- <speaker id="6" type="a">
  - <plist>
      <p>I will ask John to answer that one.</p>
    </plist>
  </speaker>
- <speaker id="5" type="q">
  - <plist>
      <p>Thanks.</p>
    </plist>
  </speaker>
- <speaker id="7" type="a">
  - <plist>
      <p>Matt, I think there are a couple of issues there. One is the products
      that have sold have tended to be less margin-rich, things like some of our
```

```
          PC sales or laptop sales and so our margin dollars are growing a little bit
          slower than we would have expected them to grow [...]
          </p>
          <p>So overall, I think that the message is that our cautious approach is really
          oriented towards continuing to invest in the business, plus seeing the mix of sales
          not be as attractive as we have had in other years.</p>
        </plist>
      </speaker>
       ...
    - <speaker id="0">
      - <plist>
          <p>And thank you for your participation in today's conference.
          This concludes the presentation. You may now disconnect. Good day.</p>
        </plist>
      </speaker>
  </section>
  </body>
</transcript>
```


A.2. Records in the database that correspond to *.xml* file

TRID	NAMECORRECTED	GVKEY	CONAME	INSTANCE	QYEAR
1079826	JOSEPH DOODY	15521	STAPLES INC	Yeah Brad, across all segments [...]	Q1 2007
1079826	JOHN MAHONEY	15521	STAPLES INC	Matt, I think there are a couple [...]	Q1 2007
1079826	RONALD SARGENT	15521	STAPLES INC	Good morning Matt. I will ask John to answer [...]	Q1 2007
1079826	DEMOS PARNEROS	15521	STAPLES INC	Yeah. I haven't seen a lot of change either [...]	Q1 2007
1079826	MICHAEL MILES	15521	STAPLES INC	Yeah Bill, it's a competitive market [...]	Q1 2007

Matt, I think there are a couple of issues there. One is the products that have sold have tended to be less margin-rich, things like some of our PC sales or laptop sales and so our margin dollars are growing a little bit slower than we would have expected them to grow.[...]

Obviously we would need to be sure that you had all the other assumptions right. We are not expecting significant improvement in our working capital performance this year in the free cash flow guidance that we have given. And I think that that's answered your question but I think we would want to sure that the other assumptions you have are also accurate before we would comment on your actual number.

Appendix B. Summary of criteria for labeling deceptive calls

We label the earnings conference call for the quarter that has its financial results later being restated as deceptive according to the following criteria:

Criterion		Description
No-threshold	NT	The restatement involves the disclosure of a material weakness (within one year, within one year before, or within one year after), or a late filing (within one year, within one year before, or within one year after), or an auditor change (within one year, within one year before, or within one year after), or a Form 8-K filing
Absolute value of cumulative bias in net income greater than 25th percentile	AS25	The same as NT and the absolute value of cumulative bias in net income over restated quarters scaled by total assets is greater than 0.06% (25th percentile of the non-zero absolute value of cumulative bias distribution)
Absolute value of cumulative bias in net income greater than 50th percentile	AS50	The same as NT and the absolute value of cumulative bias in net income over restated quarters scaled by total assets is greater than 0.21% (50th percentile of the non-zero absolute value of cumulative bias distribution)
Positive value of cumulative bias in net income greater than 25th percentile	PS25	The same as NT and the positive cumulative bias in net income over restated quarters scaled by total assets is greater than 0.06% (25th percentile of the non-zero positive cumulative bias distribution)
Positive value of cumulative bias in net income greater than 50th percentile	PS50	The same as NT and the positive cumulative bias in net income over restated quarters scaled by total assets is greater than 0.24% (25th percentile of the non-zero positive cumulative bias distribution)

References

Adams, S. H., and J. P. Jarvis. 'Indicators of veracity and deception: an analysis of written statements made to police.' *Speech, Language and the Law* 13(1) (2006): 1–22.

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): 1259–1294.

Bachenko, J., E. Fitzpatrick, and M. Schonwetter. 'Verification and implementation of language-based deception indicators in civil and criminal narratives.' *Proceedings of the 22nd International Conference on Computational Linguistics* (2008): 41–48.

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.

Bond, G. D., and A. Y. Lee. 'Language of lies in prison: linguistic classification of prisoners' truthful and deceptive natural language.' *Applied Cognitive Psychology* 19(3) (2005): 313–329.

Canty, A., and B. Ripley. boot: Bootstrap R (S-Plus) Functions. R Foundation for Statistical Computing, 2009.

Core, J. E., W. Guay, and D. F. Larcker. 'The power of the pen and executive compensation.' *Journal of Financial Economics* 88 (2008): 1–25.

Correia, M. M. 'Political Connections, SEC Enforcement and Accounting Quality.' Unpublished paper. SSRN eLibrary, 2009. Available at http://ssrn.com/paper=1458478.

Dahl, D. B. xtable: Export tables to LaTeX or HTML, 2009.

Das, S. R., and M. Y. Chen. 'Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web.' *Management Science* 53 (2007): 1375–1388.

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

Davison, A. C., and D. V. Hinkley. Bootstrap Methods and Their Application (Cambridge Series in Statistical and Probabilistic Mathematics , No 1). Cambridge University Press, 1997.

Dechow, P. M., and I. D. Dichev. 'The Quality of Accruals and Earnings: The Role of Accrual Estimation Errors.' *The Accounting Review* 77 (2002): 35–59.

Dechow, P. M., W. Ge, C. R. Larson, and R. G. Sloan. 'Predicting Material Accounting Misstatements.' *Contemporary Accounting Research, Forthcoming* (2010).

Demers, E. A., and C. Vega. 'Soft Information in Earnings Announcements: News or Noise?' Unpublished paper. SSRN eLibrary, 2010. Available at http://ssrn.com/paper=1152326.

DePaulo, B. M., J. J. Lindsay, B. E. Malone, L. Muhlenbruck, K. Charlton, and H. Cooper. 'Cues to Deception.' *Psychological Bulletin* 129(1) (2003): 74–118.

Efron, B., and R. J. Tibshirani. An Introduction to the Bootstrap. Chapman & Hall/CRC, 1994.

Fawcett, T. 'An introduction to ROC analysis.', 2006.

Feinerer, I. tm: Text Mining Package, 2010. R package version 0.5-3.

Feinerer, I., K. Hornik, and D. Meyer. 'Text Mining Infrastructure in R.' *Journal of Statistical Software* 25 (2008).

Friedman, J., T. Hastie, and R. Tibshirani. 'Regularization Paths for Generalized Linear Models via Coordinate Descent.' *Journal of Statistical Software* 33 (2009).

Hastie, T., R. Tibshirani, and J. Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Corrected ed. Springer, 2003.

Henry, E., and A. J. Leone. 'Measuring Qualitative Information in Capital Markets Research.' Unpublished paper. SSRN eLibrary, 2009. Available at http://ssrn.com/paper=1470807.

Hobson, J. L., W. J. Mayew, and M. Venkatachalam. 'Analyzing Speech to Detect Financial Misreporting.' Unpublished paper. SSRN eLibrary, 2010. Available at http://ssrn.com/paper=1531871.

Hribar, P., and D. W. Collins. 'Errors in Estimating Accruals: Implications for Empirical Research.' *Journal of Accounting Research* 40(1) (2002): 105–134.

Jones, J. J. 'Earnings Management During Import Relief Investigations.' *Journal of Accounting Research* 29 (1991): 193–228.

Kasznik, R. 'On the Association between Voluntary Disclosure and Earnings Management.' *Journal of Accounting Research* 37 (1999): 57–81.

Knapp, M. L., R. P. Hart, and H. S. Dennis. 'An exploration of deception as a communication construct.' *Human Communication Research* 1(1) (1974): 15–29.

Kothari, S., X. Li, and J. E. 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.' Unpublished paper. SSRN eLibrary, 2008. Available at http://ssrn.com/paper=1113337.

Li, F. 'Do Stock Market Investors Understand the Risk Sentiment of Corporate Annual Reports?' Unpublished paper. SSRN eLibrary, 2006. Available at http://ssrn.com/paper=898181.

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

Li, F. 'The Information Content of Forward-Looking Statements in Corporate Filings: A Naive Bayesian Machine Learning Approach.' *Journal of Accounting Research, Forthcoming* (2010).

Loughran, T., B. McDonald, and H. Yun. 'A Wolf in Sheeps Clothing: The Use of Ethics-Related Terms in 10-K Reports.' *Journal of Business Ethics* 89 (2009): 39–49.

McNichols, M. F. 'Research design issues in earnings management studies.' *Journal of Accounting and Public Policy* 19 (2000): 313 – 345.

Newman, M. L., J. W. Pennebaker, D. S. Berry, and J. M. Richards. 'Lying Words: Predicting Deception from Linguistic Styles.' *Personality and Social Psychology Bulletin* 29 (2003): 665–675.

Palmrose, Z.-V., V. Richardson, and S. Scholz. 'Determinants of market reactions to restatement announcements.' *Journal of Accounting and Economics* 37 (2004): 59–89.

Pennebaker, J. W., C. K. Chung, M. Ireland, A. Gonzales, and R. J. Booth. The Development and Psychometric Properties of LIWC2007, 2007.

Plumlee, M. A., and T. L. Yohn. 'Restatements: Investor Response and Firm Reporting Choices.' Unpublished paper. SSRN eLibrary, 2008. Available at http://ssrn.com/paper=1186254.

R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2005.

Sing, T., O. Sander, N. Beerenwinkel, and T. Lengauer. ROCR: visualizing classifier performance in R, 2005.

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

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

Vrij, A. Detecting Lies and Deceit: Pitfalls and Opportunities. 2nd ed. Wiley, 2008.

Witten, I. H., and E. Frank. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann Series in Data Management Systems, 2nd ed. Morgan Kaufmann, 2005.

Table 1. Definitions of word-based variables

This table presents definitions of the word-based variables that we use to estimate classification models for deceptive instances. Panel A outlines variable computation and predicted signs. The predicted sign is the hypothesized association with the likelihood of deception. LIWC is the Linguistic Inquiry and Word Count psychosocial dictionary by James W. Pennebaker, Roger J. Booth, and Martha E. Francis (Pennebaker et al. [2007]). Panel B lists our self-constructed word categories and individual words that these categories include.

Panel A: Variables, computation, and predicted signs			
Category	Abbreviation	Sign	Calculation
Word count	wc	+/-	Number of words ignoring articles (a, an, the). Prior research: Newman et al. [2003], Vrij [2008]
			References
1st person singular pronouns	I	-	LIWC category "I": I, me, mine, etc. Simple count dividend by the number of words ignoring articles (wc) and multiplied by the median wc in the sample. Prior research: Bachenko et al. [2008], Bond and Lee [2005], DePaulo et al. [2003], Newman et al. [2003], Knapp et al. [1974], Newman et al. [2003], Vrij [2008]
1st person plural pronouns	we	+	LIWC category "we": we, us, our, etc. Simple count dividend by the number of words ignoring articles (wc) and multiplied by the median wc in the sample. Prior research: Bachenko et al. [2008], Bond and Lee [2005], DePaulo et al. [2003], Newman et al. [2003], Knapp et al. [1974], Newman et al. [2003], Vrij [2008]
3rd person plural pronouns	they	+/-	LIWC category "they": they, their, they'd, etc. Simple count dividend by the number of words ignoring articles (wc) and multiplied by the median wc in the sample. Prior research: Vrij [2008], Knapp et al. [1974], Newman et al. [2003]

Category	Abbreviation	Sign	Calculation
Impersonal pronouns	ipron	+/-	LIWC category "ipron": it, anyone*, nobod*, etc. Simple count dividend by the number of words ignoring articles (wc) and multiplied by the median wc in the sample. Prior research: DePaulo et al. [2003], Knapp et al. [1974], Vrij [2008]
Reference to general knowledge	genknlref	+/-	Self-constructed category: you know, investors well know, others know well, etc. For the complete list see Panel B. Simple count dividend by the the number of words ignoring articles (wc) and multiplied by the median wc in the sample.
			Positives/Negatives
Assent	assent	-	LIWC category "assent": agree, OK, yes, etc. Simple count dividend by the number of words ignoring articles (wc) and multiplied by the median wc in the sample. Prior research: Vrij [2008]
Non-extreme positive emotions	posemone	-	Modified LIWC category "posemo": love, nice, accept, etc. These LIWC category excludes extreme positive emotions words which are listed in the Panel B. Simple count dividend by the number of words ignoring articles (wc) and multiplied by the median wc in the sample. Prior research: Newman et al. [2003], Vrij [2008]
Extreme positive emotions	posemoextr	+/-	Self-constructed category: fantastic, great, definitely, etc. For the complete list see Panel B. Simple count dividend by the number of words ignoring articles (wc) and multiplied by the median wc in the sample. Prior research: Newman et al. [2003], Vrij [2008]
Negations	negate	+	LIWC category "negate": no, not, never, etc. Simple count dividend by the number of words ignoring articles (wc) and multiplied by the median wc in the sample. Prior research: Adams and Jarvis [2006], Bachenko et al. [2008], Newman et al. [2003], Vrij [2008]

Category	Abbreviation	Sign	Calculation
Anxiety	anx	+	LIWC category "anx": worried, fearful, nervous, etc. Simple count dividend by the number of words ignoring articles (wc) and multiplied by the median wc in the sample. Prior research: Bachenko et al. [2008], Bond and Lee [2005], Knapp et al. [1974], Newman et al. [2003], Vrij [2008]
Anger	anger	+	LIWC category "anger": hate, kill, annoyed, etc. Simple count dividend by the number of words ignoring articles (wc) and multiplied by the median wc in the sample. Prior research: Bachenko et al. [2008], Bond and Lee [2005], Newman et al. [2003], Vrij [2008]
Swear words	swear	+	LIWC category "swear": screw*, hell, etc. Simple count dividend by the number of words ignoring articles (wc) and multiplied by the median wc in the sample. Prior research: Bachenko et al. [2008], DePaulo et al. [2003], Vrij [2008]
Extreme negative emotions	negemoextr	+	Self-constructed category: absurd, adverse, awful, etc. For the complete list see Panel B. Simple count dividend by the number of words ignoring articles (wc) and multiplied by the median wc in the sample. Prior research: Newman et al. [2003], Vrij [2008]

Category	Abbreviation	Sign	Calculation
			Cognitive process
Certainty	certain	-	LIWC category "certain": always, never, etc. Simple count dividend by the number of words ignoring articles (wc) and multiplied by the median wc in the sample. Prior research: Bond and Lee [2005], Knapp et al. [1974], Newman et al. [2003], Vrij [2008]
Tentative	tentat	+	LIWC category "tentat": maybe, perhaps, guess, etc. Simple count dividend by the number of words ignoring articles (wc) and multiplied by the median wc in the sample. Prior research: Adams and Jarvis [2006], Bond and Lee [2005], DePaulo et al. [2003], Knapp et al. [1974], Newman et al. [2003], Vrij [2008]
			Other cues
Hesitations	hesit	+/-	Self-constructed category on the basis of LIWC category "filler": ah, um, uhm, etc. For the complete list see Panel B. Simple count dividend by the number of words ignoring articles (wc) and multiplied by the median wc in the sample. Prior research: Vrij [2008]
Shareholders value	shvalue	+/-	Self-constructed category: shareholder well-being, value for our shareholders, value for shareholders, etc. For the complete list see Panel B. Simple count dividend by the number of words ignoring articles (wc) and multiplied by the median wc in the sample.
Value creation	value	+/-	Self-constructed category: value creation, unlocks value, improve value, etc. For the complete list see Panel B. Simple count dividend by the number of words ignoring articles (wc) and multiplied by the median wc in the sample.

Panel B: Self-constructed word categories	
Reference to general knowledge	you know, you guys know, you folks know, you well know, you long know, you would agree, everybody knows, everybody well knows, everybody long knows, everybody would agree, everyone knows, everyone well knows, everyone long knows, everyone would agree, others know, others well know, others long know, others would agree, they know, they well know, they long know, they would agree, investors know, investors well know, investors long know, investors would agree, shareholders know, shareholders well know, shareholders long know, shareholders would agree, stockholders know, stockholders well know, stockholders long know, stockholders would agree
Shareholders value	shareholder value, shareholder welfare, shareholder well-being, value for our shareholders, value for shareholders, stockholder value, stockholder welfare, stockholder well-being, value for our stockholders, value for stockholder, investor value, investor welfare, investor well-being, value for our investors, value for investors
Value creation	value creation, create value, creates value, creating value, value unlock, unlock value, unlocks value, unlocking value, value improvement, improve value, improves value, improving value, value increase, increase value, increases value, increasing value, value delivery, deliver value, delivers value, delivering value, value enhancement, enhance value, enhances value, enhancing value, value expansion , expand value, expands value, expanding value
hesit	ah, blah, eh, ehh, ehhh, hm, hmm, hmmm, huh, huhh, mm, mmm, mmmm, oh, sigh, uh, uhh, uhhs, uhm, uhmm, um, umm, zz, zzz

Extreme negative emotions	abominable, abortive, absurd, advers*, ambitious, annihilating, annihilative, atrocious, awful, badly, baffling, barbarous, bias, breach, brokenhearted, brutal*, calamitous, careless*, catchy, challenging, cockeyed, coerce, crafty, craz*, cruel*, crushed, cunning, curious, danger*, daunting, daze*, defect*, degrad*, demanding, demeaning, depress*, derisory, despair*, desperat*, despicable, destroy*, devastat*, devil*, difficult*, dire, direful, disastrous, disgraceful, dodgy, dread*, exasperating, exorbitant, extortionate, fail*, farcical, farfetched, fatal*, fateful, fault*, fearful*, fearsome, fierce, finished, fright*, frustrat*, funny, grave*, griev*, guileful, hard, harebrained, harm, harmed, harmful*, harming, harms, heartbreak*, heartbroke*, heartless*, heartrending, heartsick, hideous, hopeless*, horr*, humbling, humiliat*, hurt*, idiot, idiotic, ignominious, ignor*, implausible, impossible, improbable, inauspicious, inconceivable, inferior* , infuriating, inglorious, insane, insecur*, intimidat*, jerk, jerked, jerks, kayoed, knavish, knocked out, knotty, KOd out, KO'd out, laughable, life-threatening, luckless*, ludicrous*, maddening, madder, maddest, maniac*, menace, mess, messy, miser*, misfortunate, mortifying, muddle, nast*, nonsensical, outrag*, overwhelm*, painf*, panic*, paranoi*, pathetic*, peculiar*, pessimis*, pickle, piti*, precarious, preconception, prejudic*, preposterous, pressur*, problem*, reek*, resent*, ridicul*, roughshod, ruin*, savage*, scandalous, scourge, serious, seriously, severe*, shake*, shaki*, shaky, shame*, shock*, silly, skeptic*, slimy, slippery, squeeze, steep, strange, stunned, stupefied, stupid*, suffer, suffered, sufferer*, suffering, suffers, sunk, terribl*, terrified, terrifies, terrify, terrifying, terror*, threat*, thwarting, ticked, tough*, tragic* , transgress, trauma*, tremendous, trick*, trigger-happy, ugl*, unbelievable, unconscionable, unconvincing, unimaginable, unimportant, unlucky, unmanageable, unspeakable, unsuccessful*, untoward, unworthy, usurious, vehement, vexing, vicious*, victim*, vile, violat*, violent*, vulnerab*, washed-up, wicked*, withering, wonky, worst, worthless* , wretched, very bad

Extreme positive emotions	amaz*, A-one, astonish*, awe-inspiring, awesome, awful, bang-up, best, bless*, brillian*, by all odds, careful*, challeng*, cherish*, confidence, confident, confidently, convinc*, crack, cracking, dandy, deadly, definite, definitely, delectabl*, delicious*, deligh*, deucedly, devilishly, dynam*, eager*, emphatically, enormous, excel*, excit*, exult, fab, fabulous*, fantastic*, first-rate, flawless*, genuinely, glori*, gorgeous*, grand, grande*, gratef*, great, groovy, hero*, huge, illustrious, immense, in spades, in truth, incredibl*, insanely, inviolable, keen*, luck, lucked, lucki*, lucks, lucky, luscious, madly, magnific*, marvellous, marvelous, neat*, nifty, outstanding, peachy, perfect*, phenomenal, potent, privileg*, rattling, redoubtable, rejoice, scrumptious*, secur*, sincer*, slap-up, smashing, solid, splend*, strong*, substantial, succeed*, success*, super, superb, superior*, suprem*, swell, terrific*, thankf*, tiptop, topnotch, treasur*, tremendous, triumph*, truly, truth*, unassailable, unbelievable, unquestionably, vast, wonderf*, wondrous, wow*, yay, yays, very good

Table 2. Descriptive statistics: exchange membership and industry composition.

This table presents exchange membership (Panel A) and industry composition (Panel B) for Compustat universe and for the three overlapping samples: the sample of CEOs (CEO), the sample of CFOs (CFO), and the sample for the model that includes discretionary accruals (DAC).

Panel A: Firms by stock exchange in 2005				
	Compustat, %	CEO, %	CFO, %	DAC, %
Non-traded Company or Security	2.42	0.27	0.31	0.29
New York Stock Exchange	27.44	45.43	47.54	45.99
NYSE Amex	6.16	1.37	1.01	0.94
OTC Bulletin Board	9.97	1.49	1.01	1.08
NASDAQ-NMS Stock Market	39.00	46.02	44.98	46.71
NYSE Arca	2.36	0.04	0.00	0.00
Other-OTC	12.64	5.37	5.15	4.99
Number of observations	8083	2549	2272	1383

Panel B: Firms by industry in 2005				
	Compustat, %	CEO, %	CFO, %	DAC, %
Mining/Construction	1.69	2.00	1.85	2.17
Food	1.53	2.04	2.16	2.75
Textiles/Print/Publish	2.90	4.63	4.49	5.93
Chemicals	1.98	2.43	2.55	3.33
Pharmaceuticals	6.34	5.88	4.97	5.35
Extractive	3.55	3.49	3.26	4.12
Durable Manufacturing	15.95	19.14	18.31	23.28
Computers	12.16	15.06	15.40	19.67
Transportation	4.69	6.08	6.43	7.81
Utilities	3.92	3.53	3.92	0.00
Retail	7.09	10.63	11.05	13.88
Financial	14.19	11.10	11.14	0.00
Insurance/RealEstate	14.49	4.71	4.67	0.00
Services	7.72	8.79	9.42	11.21
Other Industries	1.80	0.47	0.40	0.51
Number of observations	8281	2549	2272	1383

Table 3. Descriptive statistics: deceptive firm-quarters by year

This table reports the frequency of deceptive firm-quarters by year. The first column is the total number of firm-quarters by year. The following three columns are counts of deceptive firm-quarters under the different criteria described in Appendix B: NT, AS25, and AS50; where N is the count of deceptive firm-quarters under a particular criterion and % is the percentage of deceptive firm-quarters in the total number of firm-quarters in a particular year.

Panel A: Deceptive firm-quarters by year for CEO sample							
		NT		AS25		AS50	
	N	N	%	N	%	N	%
2003	1108	227	20.49	80	7.22	64	5.78
2004	4051	828	20.44	357	8.81	230	5.68
2005	4516	720	15.94	429	9.50	291	6.44
2006	5480	417	7.61	271	4.95	211	3.85
2007	1422	60	4.22	34	2.39	29	2.04
Total	16577	2252	13.59	1171	7.06	825	4.98
Panel B: Deceptive firm-quarters by year for CFO sample							
		NT		AS25		AS50	
	N	N	%	N	%	N	%
2003	1021	207	20.27	84	8.23	64	6.27
2004	3593	748	20.82	322	8.96	204	5.68
2005	3959	649	16.39	392	9.90	252	6.37
2006	4652	401	8.62	243	5.22	178	3.83
2007	1237	48	3.88	28	2.26	22	1.78
Total	14462	2053	14.20	1069	7.39	720	4.98
Panel C: Deceptive firm-quarters by year for DAC sample							
		NT		AS25		AS50	
	N	N	%	N	%	N	%
2003	392	90	22.96	46	11.73	39	9.95
2004	1369	333	24.32	171	12.49	121	8.84
2005	1527	323	21.15	217	14.21	144	9.43
2006	1595	195	12.23	131	8.21	105	6.58
2007	298	15	5.03	10	3.36	9	3.02
Total	5181	956	18.45	575	11.10	418	8.07

Table 4. Descriptive statistics: explanatory variables for the sample of CEOs (CFOs)

This table reports descriptive statistics for the variables that we include in our binomial logistic models. Panel A contains descriptive statistics for the sample of CEOs; Panel B for the sample of CFOs. Variables are winsorized at 1- and 99- percentiles. The word categories are defined in Table 1.

Panel A: Descriptive statistics for CEO sample (N = 16577)							
	Mean	Std dev	25th	50th	75th	Min	Max
wc	1811.46	1129.40	931.00	1611.00	2480.00	192.00	5315.16
References							
I	29.25	12.63	20.15	28.01	36.94	4.96	67.03
we	84.01	20.71	69.81	82.96	97.45	37.67	137.51
they	11.89	8.07	6.14	10.40	16.07	0.00	39.66
ipron	121.24	19.32	108.30	121.00	133.73	74.89	172.23
genknlref	5.26	7.14	0.68	2.67	6.71	0.00	37.60
Positives/Negatives							
assent	5.91	4.40	2.87	5.04	7.95	0.00	22.38
posemone	52.34	13.50	43.06	51.11	60.28	24.20	93.63
posemoextr	8.79	5.31	5.08	8.00	11.56	0.00	26.97
negate	22.53	8.87	16.29	21.73	27.92	4.54	48.91
anx	1.46	1.84	0.00	0.93	2.19	0.00	9.24
anger	1.37	1.71	0.00	0.91	2.08	0.00	8.29
swear	0.10	0.36	0.00	0.00	0.00	0.00	2.23
negemoextr	3.38	2.79	1.34	2.93	4.91	0.00	13.03
Cognitive mechanism							
certain	23.15	7.84	17.76	22.45	27.83	6.53	46.59
tentat	50.64	13.72	41.27	49.86	59.30	19.39	88.00
Other cues							
hesit	0.19	0.54	0.00	0.00	0.00	0.00	3.24
shvalue	0.05	0.25	0.00	0.00	0.00	0.00	1.77
value	0.04	0.20	0.00	0.00	0.00	0.00	1.45

Panel B: Descriptive statistics for CFO sample (N = 14462)							
	Mean	Std dev	25th	50th	75th	Min	Max
wc	987.82	750.02	438.25	777.50	1303.00	161.00	3780.51
References							
I	12.59	6.46	7.96	11.95	16.37	0.00	32.59
we	39.30	10.80	31.83	38.91	46.34	13.95	67.33
they	3.46	3.28	1.06	2.74	4.94	0.00	16.09
ipron	57.60	10.94	50.34	57.43	64.60	31.57	87.39
genknlref	2.55	3.69	0.00	1.16	3.42	0.00	18.70
Positives/Negatives							
assent	4.32	3.45	1.94	3.59	5.86	0.00	17.34
posemone	22.52	7.49	17.43	22.07	27.00	6.24	44.37
posemoextr	2.70	2.33	0.89	2.35	4.01	0.00	10.24
negate	10.37	5.00	6.96	9.84	13.22	0.00	26.21
anx	0.50	0.90	0.00	0.00	0.76	0.00	4.45
anger	0.35	0.71	0.00	0.00	0.47	0.00	3.63
swear	0.03	0.15	0.00	0.00	0.00	0.00	1.13
negemoextr	1.14	1.38	0.00	0.76	1.81	0.00	6.37
Cognitive mechanism							
certain	10.35	4.66	7.19	9.97	13.11	0.00	24.32
tentat	23.56	7.57	18.35	23.22	28.23	6.64	44.64
Other cues							
hesit	0.10	0.37	0.00	0.00	0.00	0.00	2.39
shvalue	0.01	0.08	0.00	0.00	0.00	0.00	0.67
value	0.00	0.04	0.00	0.00	0.00	0.00	0.39

Figure 1. ROC curve for the sample of CEOs of financial firms

The ROC curve is produced by the threshold averaging over 100 cross-validation runs. Specifically, each cutoff (threshold) for the probability of deception corresponds to a point on the ROC curve. For each cutoff, we average the true positive rate and the false positive rate for the 100 out-of-sample classifications, which we obtain by 5-fold cross-validation repeated 20 times. The diagonal is the ROC curve for a random classifier. The bars on the ROC curve show the 95% confidence intervals at the cutoffs for the probability of deception equal to 0.25, 0.50, and 0.75.

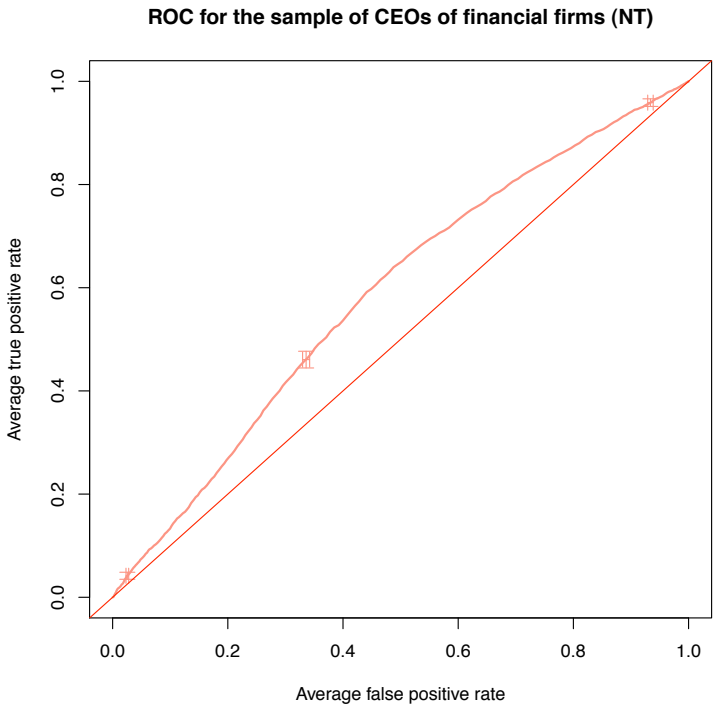

Table 5. Results for financial firms: the no-threshold criterion

This table reports classification results of the logistic models that use all word-based variables defined in Table 1 for CEOs (Panel A) and CFOs (Panel B) to predict deceptive instances under no-threshold criterion (Appendix B). The first half of the table presents means over 100 cross-validation runs of out-of-sample performance measures: AUC (the area under ROC) in percentages, precision (the percentage of actual deceptive instances among those classified by the algorithm as deceptive), and accuracy (the percentage of correctly classified instances). Here, 100 cross-validation runs is a result of 5-fold cross-validation repeated 20 times. Here, "t-test vs. 50%" is the value of the t-statistic testing the null hypothesis of the mean AUC being equal to 50% which is the AUC of a random classifier. The second half of the table reports factors by which the odds of a deceptive instance is multiplied if the number or words in a particular category increases by 1% of the median instance length. Specifically, for CEOs we report $e^{15\beta}$ (i.e., 1% of 1,526), and for CFOs we report $e^{7\beta}$ (i.e., 1% of 708). For the word count $wc^{\ddagger}$, we present the effect of increasing the instance length by the the median instance length (i.e., $e^{1,526\beta}$ for CEOs and $e^{708\beta}$ for CFOs). Only factors for coefficients significant at 10% (two-tailed) are shown; estimates of intercepts are omitted. We perform 1000 stratified bootstrap replications to compute percentile confidence intervals. Explanatory variables are winsorized at 1- and 99- percentile to mitigate the effect of outliers. Here, $*$, $**$, and $***$ denote correspondingly factors for coefficients significant at 10%, 5%, and 1% significance level.

		Panel A: CEO	Panel B: CFO
Sample composition			
Total firm-quarters		1533	1265
Deceptive firm-quarters		160	130
Deceptive firm-quarters(%)		10.44	10.28
AUC, Precision, and Accuracy in %			
AUC (t-test vs. 50 %)		58.93 (20.26)	53.56 (6.37)
Precision		13.76	11.57
Accuracy		64.27	62.61
Logistic regression			
Word count			
$wc^{\ddagger}$	$+/-$		1.24**
References			
ipron	$+/-$		0.90*
genknlref	$+/-$	1.69***	
Positives/Negatives			
assent	$-$	0.55*	
negemoextr	$+$		3.48**
Cognitive mechanism			
certain	$-$	1.72***	
tentat	$+$	0.77**	0.80**

Table 6. AUC, Precision, and Accuracy for the samples of CEOs (CFOs) for the NT, AS25, and AS50 criteria

This table reports classification performance of the logistic models that use all word-based variables defined in Table 1 for CEOs (Panel A) and CFOs (Panel B) to predict deceptive instances under NT, AS25, and AS50 criteria (Appendix B). We compute means over 100 cross-validation runs of out-of-sample performance measures: AUC (the area under ROC) in percentages, precision (the percentage of actual deceptive instances among those classified by the algorithm as deceptive), and accuracy (the percentage of correctly classified instances). Here, 100 cross-validation runs is a result of 10-fold cross-validation repeated 10 times. Here, "t-test vs. 50%" is the value of the t-statistic testing the null hypothesis of the mean AUC being equal to 50% which is an AUC of a random classifier. Explanatory variables are winsorized at 1- and 99- percentile to mitigate the effect of outliers.

Panel A: CEO sample			
Sample composition			
	NT	AS25	AS50
Total firm-quarters	16577	16577	16577
Deceptive firm-quarters	2252	1171	825
Deceptive firm-quarters(%)	13.59	7.06	4.98
AUC, Precision, and Accuracy in %			
AUC (t-test vs. 50 %)	56.37 (32.05)	54.06 (15.84)	56.29 (21.22)
Precision	16.73	7.90	5.84
Accuracy	65.03	53.10	53.07
Panel B: CFO			
Sample composition			
	NT	AS25	AS50
Total firm-quarters	14462	14462	14462
Deceptive firm-quarters	2053	1069	720
Deceptive firm-quarters(%)	14.2	7.39	4.98
AUC, Precision, and Accuracy in %			
AUC (t-test vs. 50 %)	56.35 (33.53)	53.67 (14.12)	54.83 (16.23)
Precision	16.80	8.41	5.70
Accuracy	58.31	59.54	52.43

Table 7. Logistic regression for the sample of CEOs (CFOs) for the NT, AS25, and AS50 criteria

This table summarizes estimation of the logistic models that use all word-based variables defined in Table 1 for CEOs (Panel A) and CFOs (Panel B) to predict deceptive instances under NT, AS25, and AS50 criteria (Appendix B). The table reports factors by which the odds of a deceptive instance is multiplied if the number or words in a particular category increases by 1% of the median instance length. Specifically, for CEOs we report $e^{16\beta}$ (i.e, 1% of 1,611), and for CFOs we report $e^{7\beta}$ (i.e., 1% of 777). For the word count $wc^{\ddagger}$, we present the effect of increasing the instance length by the median instance length (i.e., $e^{1{,}611\beta}$ for CEOs and $e^{777\beta}$ for CFOs); whereas for $swear^{\dagger}$, $hesit^{\dagger}$, $shvalue^{\dagger}$ and $value^{\dagger}$ the effect of increasing in the number of words in the corresponding category by one word (i.e., e^{β}). Only factors for coefficients significant at 10% (two-tailed) are shown; estimates of intercepts are omitted. We perform 1000 stratified bootstrap replications to compute percentile confidence intervals. Explanatory variables are winsorized at 1- and 99- percentile to mitigate the effect of outliers. Here, $*$, $**$, and $***$ denote correspondingly factors for coefficients significant at 10%, 5%, and 1% significance level.

Panel A: CEO sample (N = 16577)				
		NT	AS25	AS50
Word count				
$wc^{\ddagger}$	+/−			1.10*
References				
I	−	0.93**		
we	+	0.94***		1.05*
they	+/−			1.24***
ipron	+/−			1.08***
genknlref	+/−	1.49***	1.23***	
Positives/Negatives				
assent	−			1.30*
posemone	−	0.93**		
posemoextr	+/−	1.23***		
anx	+	0.67*	0.47**	0.43**
$swear^{\dagger}$	+	0.86**		0.78**
negemoextr	+	0.76**	0.67**	0.63*
Cognitive mechanism				
certain	−			0.85**
Other cues				
$hesit^{\dagger}$	+/−			0.88*
$shvalue^{\dagger}$	+/−	0.84*		
$value^{\dagger}$	+/−	0.75**	0.65***	0.67**

Panel B: CFO sample (N = 14462)				
		NT	AS25	AS50
Word count				
wc‡	+/−	1.08***	1.11***	
References				
they	+/−	0.87**	0.84**	
genknlref	+/−	1.39***	1.18**	1.14**
Positives/Negatives				
assent	−	1.11**		
posemone	−	0.92***	0.90***	0.82***
anx	+	0.67**		
Cognitive mechanism				
certain	−	1.08**	1.08*	
tentat	+	1.04*		
Other cues				
shvalue†	+/−	0.40**		
value†	+/−			0.12*

Table 8. Descriptive statistics: discretionary accruals model

This table reports descriptive statistics for the discretionary accruals model. The unrestated quarterly data is from Computstat Point In Time. For a number of firms, some balance sheet variables, e.g. Gross PP&E, are reported only in the fourth quarter. To fill in missing values for balance sheet variables, we exploit their persistence property and extrapolate them three quarters forward and back (take averages if possible). If the financial results for the same fiscal quarter correspond to several datadates, we take the average of final variables. If cash flow from operations item is missing we compute cash flow from operations following balance sheet method: $CFO_t = IBQ_t - ((ACTQ_t - ACTQ_{t-1}) - (LCTQ_t - LCTQ_{t-1}) - (CHEQ_t - CHEQ_{t-1}) + (DLCQ_t - DLCQ_{t-1}) - DPQ_t)$. All variables are winsorized at 1-, 99- percentiles. The coefficients from total accruals regressions are not winsorized. To estimate the discretionary part of total accruals, we replicate Kasznik [1999], and estimate the following regression by 2-digit SIC code and fiscal quarter:

$$TAC_{it} = \beta_0 + \beta_1(\Delta Sales_{it} - \Delta AR_{it}) + \beta_2 PPE_{it} + \beta_3 \Delta CFO_{it} + \epsilon_{it}$$

Panel A: Definition of variables for the discretionary accruals model		
Category	Abbreviation	Calculation
Average total assets	mAT	$\frac{ATQ_t + ATQ_{t-1}}{2}$; or ATQ_{t-1} if ATQ_t is missing; or ATQ_t if ATQ_{t-1} is missing
Operating cash flow	CFO	$OANCFQ_t$; or if missing calculated by balance sheet method
Total accruals	TAC	$IBCQ_t - (CFO_t - XIDOCQ_t)$ as in Hribar and Collins [2002]; or if missing calculated as $NIQ_t - CFO_t$
Change in sales adjusted for change in receivables	dSALEmdAR	$(SALEQ_t - SALEQ_{t-1}) - (RECTQ_t - RECTQ_{t-1})$
Gross PP&E	PPE	$PPEGTQ_t$
Change in CFO	dCFO	$CFO_t - CFO_{t-1}$
nTAC, ndSALEmdAR, nPPE, ndCFO are corresponding variables scaled by mAT		

Panel B: Model for prediction of deceptive instances		
Category	Abbreviation	Calculation
Discretionary accruals	nDAC	Defined as the difference between total accruals and non-discretionary accruals estimated as fitted value from the regression above
Actual issuance	CAPMKT	An indicator variable coded 1 if the firm issued securities or long-term debt ($SSTKQ > 0 or DLTISQ > 0$) and 0 otherwise
Market capitalization	MCAP	$CSHOQ_t \cdot PRCCQ_t$
Free cash flow	FCF	$CFO_t - CAPXQ\_Mean_t$ where we compute $CAPXQ\_Mean_t$ over 12 quarters requiring at least three non-missing observations
Growth in cash sales	GROWTH	$(SALEQ_t - \Delta AR_t)/(SALEQ_{t-1} - \Delta AR_{t-1}) - 1$
nMCAP, nFCF are corresponding variables scaled by mAT		

Descriptive statistics for the discretionary accruals model (N = 5181)							
	Mean	Std dev	25th	50th	75th	Min	Max
nMCAP	1.73	1.44	0.80	1.29	2.23	0.11	8.00
nFCF	0.03	0.07	-0.01	0.03	0.07	-0.26	0.24
GROWTH	0.05	0.23	-0.05	0.02	0.11	-0.56	1.21
nDAC	0.03	0.09	-0.02	0.02	0.07	-0.23	0.32

Actual issuance			
	N	Percentage of 0	Percentage of 1
Actual issuance	5181	3.57	96.43

Table 9. AUC: comparison of models based on verbal cues or/and financial variables for the NT, AS25, and AS50 criteria

This table presents AUC (the area under ROC) in percentages for the models that predict deceptive instances using verbal cues or/and financial variables. Here, NT (Panel A), AS25 (Panel B), and AS50 (Panel C) denote criteria for labeling instances as deceptive (Appendix B). We perform 10-fold cross-validation repeated 10 times which provides us with 100 out-of-sample performance measures. Six binomial logistic models with different set of variables - word-based CEO variables (CEO), word-based CFO variables (CFO), financial variables (DAC), financial variables and word-based CEO variables (DAC+CEO), financial variables and word-based CFO variables (DAC+CFO), financial variables and word-based CEO/CFO variables (DAC+CEO+CFO) - are estimated and tested on the same split of data. Here, "t-test vs. 50%" is the value of the t-statistic testing the null hypothesis of the mean AUC being equal to 50% which is an AUC of a random classifier. Values in the columns labeled as "CFO", "DAC", "DAC+CEO", "DAC+CFO", "DAC+CEO+CFO" are paired t-statistics that test the null that the paired difference in means between the performance measure for the model in the row and the performance measure for the model in the column is zero[28]. Explanatory variables are winsorized at 1- and 99- percentile to mitigate the effect of outliers.

Sample composition			
	NT	AS25	AS50
Total firm-quarters	5181	5181	5181
Deceptive firm-quarters	956	575	418
Deceptive firm-quarters(%)	18.45	11.1	8.07

Panel A: NT							
	mean %	t-test vs 50%	CFO	DAC	DAC +CEO	DAC +CFO	DAC +CEO+CFO
CEO	55.92	18.65	0.80	6.40	-3.59	0.47	-7.30
CFO	55.59	18.10		5.95	-2.14	-1.10	-8.84
DAC	52.91	9.01			-9.21	-7.67	-12.42
DAC +CEO	56.47	19.52				2.02	-6.45
DAC +CFO	55.73	19.17					-10.08
DAC +CEO+CFO	57.86	24.30					

[28]For example, in Panel A, the value 0.80 in the CFO column is the value of the paired t-statistic that tests the null hypothesis that the paired difference in means between AUC for the model estimated using word-based CEO variables and AUC for the model estimated using word-based CFO variables is zero

Panel B: AS25							
	mean %	t-test vs 50%	CFO	DAC	DAC +CEO	DAC +CFO	DAC +CEO+CFO
CEO	53.43	9.67	-2.18	-3.92	-9.72	-5.10	-9.47
CFO	54.61	12.28		-1.39	-1.58	-6.19	-5.80
DAC	55.36	16.02			-0.37	-1.94	-2.98
DAC +CEO	55.50	15.56				-1.25	-3.76
DAC +CFO	56.08	18.25					-1.90
DAC +CEO+CFO	56.59	19.73					
Panel C: AS50							
	mean %	t-test vs 50%	CFO	DAC	DAC +CEO	DAC +CFO	DAC +CEO+CFO
CEO	55.61	12.56	-0.65	1.80	-5.03	-1.46	-6.18
CFO	56.09	12.47		2.25	-0.95	-2.12	-4.67
DAC	54.39	9.29			-4.78	-3.92	-6.89
DAC +CEO	56.82	15.45				0.31	-3.55
DAC +CFO	56.62	15.08					-4.66
DAC +CEO+CFO	58.21	21.14					

Table 10. AUC, Precision, and Accuracy for the samples of CEOs (CFOs) for the PS25 and PS50 criteria

This table reports classification performance of the logistic models that use all word-based variables defined in Table 1 for CEOs (Panel A) and CFOs (Panel B) to predict deceptive instances under PS25 and PS50 criteria (Appendix B). We compute means over 100 cross-validation runs of out-of-sample performance measures: AUC (the area under ROC) in percentages, precision (the percentage of actual deceptive instances among those classified by the algorithm as deceptive), and accuracy (the percentage of correctly classified instances). Here, 100 cross-validation runs is a result of 10-fold cross-validation repeated 10 times. Here, "t-test vs. 50%" is the value of the t-statistic testing the null hypothesis of the mean AUC being equal to 50% which is the AUC of a random classifier. Explanatory variables are winsorized at 1- and 99- percentile to mitigate the effect of outliers.

Panel A: CEO sample		
Sample composition		
	PS25	PS50
Total firm-quarters	16577	16577
Deceptive firm-quarters	889	609
Deceptive firm-quarters(%)	5.36	3.67
AUC, Precision, and Accuracy %		
AUC (t-test vs. 50 %)	53.96 (14.01)	55.74 (18.72)
Precision	6.00	4.21
Accuracy	54.65	52.42
Panel B: CFO		
Sample composition		
	PS25	PS50
Total firm-quarters	14462	14462
Deceptive firm-quarters	813	539
Deceptive firm-quarters(%)	5.62	3.73
AUC, Precision, and Accuracy %		
AUC (t-test vs. 50 %)	53.62 (11.40)	53.67 (8.81)
Precision	6.38	4.20
Accuracy	59.68	56.24

Table 11. AUC, Precision, and Accuracy for adjusted vs. unadjusted word categories for the NT, AS25, and AS50 criteria

This table reports out-of-sample classification performance for the logistic models that predict deceptive instances using adjusted and unadjusted word categories listed in Table 1 for CEOs (Panel A) and CFOs (Panel B). The word categories are adjusted by subtracting the mean over previous non-deceptive instances for every individual in the sample (requiring at least two quarters). It reports AUC (the area under ROC) in percentages, precision (the percentage of actual deceptive instances among those classified by the algorithm as deceptive), and accuracy (the percentage of correctly classified instances). Here, NT, AS25, and AS50 denote criteria for labeling instances as deceptive (Appendix B). We perform 10-fold cross-validation repeated 10 times which provides us with 100 out-of-sample performance measures. Two binomial logistic models with different set of variables - adjusted and unadjusted word categories - are estimated and tested on the same split of data. Here, "t-test vs. 50%" is the value of the t-statistic testing the null hypothesis of the mean AUC being equal to 50% which is the AUC of a random classifier. The column "t-test" reports paired t-statistics that test the null that the paired difference in means between the performance measure for the model with unadjusted and adjusted categories is zero. Explanatory variables are winsorized at 1- and 99- percentile to mitigate the effect of outliers.

Panel A: CEO			
Sample composition			
	NT	AS25	AS50
Total firm-quarters	9500	10440	10647
Deceptive firm-quarters	198	321	243
Deceptive firm-quarters(%)	2.08	3.07	2.28
NT			
	Unadjusted	Adjusted	t-test
AUC (t-test vs. 50%)	60.20(15.30)	56.99(11.52)	4.27
Precision	2.96	2.74	3.03
Accuracy	61.68	61.22	2.34
AS25			
	Unadjusted	Adjusted	t-test
AUC (t-test vs. 50%)	56.43(13.08)	58.34(17.44)	-2.62
Precision	3.79	4.08	-3.11
Accuracy	56.92	59.81	-12.91
AS50			
	Unadjusted	Adjusted	t-test
AUC (t-test vs. 50%)	56.85(11.69)	55.55(9.69)	1.89
Precision	2.67	2.75	-1.08
Accuracy	57.28	58.41	-5.71

Panel B: CFO			
Sample composition			
	NT	AS25	AS50
Total firm-quarters	7981	8792	8977
Deceptive firm-quarters	152	256	168
Deceptive firm-quarters(%)	1.9	2.91	1.87
	NT		
	Unadjusted	Adjusted	t-test
AUC (t-test vs. 50%)	50.50(0.73)	56.14(7.42)	-5.70
Precision	2.00	2.26	-3.26
Accuracy	54.37	61.86	-28.40
	AS25		
	Unadjusted	Adjusted	t-test
AUC (t-test vs. 50%)	54.66(8.70)	57.09(12.09)	-3.63
Precision	3.37	3.78	-5.40
Accuracy	56.41	60.32	-20.04
	AS50		
	Unadjusted	Adjusted	t-test
AUC (t-test vs. 50%)	57.53(12.46)	60.22(14.02)	-4.01
Precision	2.17	2.50	-4.59
Accuracy	56.73	63.40	-36.94

Table 12. Logistic regression for the samples of CEOs (CFOs) with adjusted word categories for the NT, AS25, and AS50 criteria

This table summarizes estimation of the logistic models that use all adjusted word categories defined in Table 1 for CEOs (Panel A) and CFOs (Panel B) to predict deceptive instances under NT, AS25, and AS50 criteria (Appendix B). The word categories are adjusted by subtracting the mean over previous non-deceptive instances for every individual in the sample (requiring at least two instances). The table reports factors by which the odds of a deceptive instance is multiplied if the number or words in a particular category increases by 1% of the median instance length. Specifically, for CEOs we report $e^{16\beta}$ (i.e, 1% of 1,611), and for CFOs we report $e^{7\beta}$ (i.e., 1% of 777). For the word count $wc^{\ddagger}$, we present the effect of increasing the instance length by the median instance length (i.e., $e^{1,611\beta}$ for CEOs and $e^{777\beta}$ for CFOs); whereas for $swear^{\dagger}$, $hesit^{\dagger}$, $shvalue^{\dagger}$ and $value^{\dagger}$ the effect of the increase in the number of words in the corresponding category by one word (i.e., e^{β}). Only factors for coefficients significant at 10% (two-tailed) are shown; estimates of intercepts are omitted. We perform 1000 stratified bootstrap replications to compute percentile confidence intervals. Explanatory variables are winsorized at 1- and 99-percentile to mitigate the effect of outliers. Here, $*$, $**$, and $***$ denote correspondingly factors for coefficients significant at 10%, 5%, and 1% significance level.

Panel A: Adjusted CEO sample				
		NT	AS25	AS50
Word count				
$wc^{\ddagger}$	+/−			1.28^{*}
References				
I	−	0.69^{***}		
they	+/−		1.53^{***}	1.31^{*}
ipron	+/−			1.13^{*}
genknlref	+/−	0.67^{**}	0.77^{*}	0.74^{**}
Positives/Negatives				
assent	−		0.63^{*}	
posemoextr	+/−	1.72^{**}		
negate	+		0.76^{**}	
$swear^{\dagger}$	+		0.73^{*}	
Cognitive mechanism				
certain	−		1.45^{***}	1.69^{***}
tentat	+	1.27^{**}		
Other cues				
$hesit^{\dagger}$	+/−	1.35^{**}		0.78^{**}
$shvalue^{\dagger}$	+/−	1.83^{**}		
$value^{\dagger}$	+/−		0.38^{***}	0.50^{**}

Panel B: Adjusted CFO sample				
		NT	AS25	AS50
References				
I	−	0.80**		
ipron	+/−	1.16**		
genknlref	+/−	0.74*	0.58***	0.50***
Positives/Negatives				
assent	−	0.67**		
posemone	−	1.24**	0.89*	0.80***
posemoextr	+/−	0.64*	0.70*	
anger	+		0.36*	
swear†	+			6.12***
negemoextr	+		2.15**	2.11*
Other cues				
value†	+/−	19.65***		0.07**